مقارنة ثلاثة إحصائيات لفحص جودة مطابقة الفقرة تحت ظروف نوع النموذج الرياضي و طول الاختبار و حجم العينة و التفاعلات بينها.

إعداد

راجي عوض الصرايرة

المشرف الأستاذ الدكتور خليل عليان

قدمت هذه الأطروحة استكمالا لمتطلبات الحصول على درجة الدكتوراه في

علم النفس التربوي / قياس و تقويم

كلية الدراسات العليا

الجامعة الأردنية

كانون ثاني، ٢٠٠٨

الجامعة الأردنية

نموذج تفويض

أنا راجي عوض مسلم الصرايرة ، أفوض الجامعة الأردنية بتزويد نسخ من أطروحتي للمكتبات أو المؤسسات أو الهيئات أو الأشخاص عند طلبها.

التوقيع:

التاريخ: ١٧/١/٢٠٠٣

قرار لجنة المناقشة

نوقشت هذه الأطروحة(مقارنة ثلاثة إحصائيات لفحص جودة مطابقة الفقرة تحت ظروف نوع النموذج الرياضي و طول الاختبار و حجم العينة و التفاعلات بينها) وأجيزت بتاريخ :٢٧/ ١٢ /٢٠٠٧

أعضاء لجنة المناقشة	التوقيع
الأستاذ الدكتور خليل محمد عليان، مشرفا أستاذ القياس و التقويم	....
الدكتور يونس محمدعبدالله اليونس،عضوا أستاذ مشارك قياس و تقويم و إحصاء	....
الدكتور عايش موسى غرايبة،عضوا أستاذ مساعد قياس و تقويم	....
الدكتور ساري سليم سواقد، عضوا أستاذ مشارك قياس و تقويم (جامعة اليرموك)	....

شكر وتقدير

يسعدني أن أتقدم بجزيل الشكر والتقدير إلى أستاذي الفاضل الذي اعتز بالتتلمذ على يديه :الأستاذ الدكتور خليل عليان الذي رعى البحث مذ كان فكرة حتى غدا على الصورة التي هو عليها.

عهدي بك – أستاذي- انك الحريص على جهد طلابك ،تقدر ظروفهم ،و أوضاعهم، فأنت ينبوع عطاء لا ينضب ،و بحر من العلم لن ينفذ أبدا ،فدم أستاذي لنا ،منك نستشرف الأمل بمستقبل زاهر ،علمتنا الدقة في العمل لنكون طلاب علم لا طلاب مؤهلات ،إن القلم ليعجز عن أن يسطر لك عبارات المدح والثناء،فأنت اجل و ارفع من شهادات حق يتفوه بها تلميذ ما زال يحبو في طريق المعرفة.

ولا يفوتني أن أتوجه بعظيم الشكر والعرفان لأساتذتي الأفاضل الدكتور يونس اليونس، والدكتور عايش غرايبة ،والدكتور ساري سواقد، الذين تفضلوا مشكورين بقبول مناقشة هذه الأطروحة،فعلى الرغم من كثرة مشاغلكم لكنكم آثرتم أن تقدموا لنا عصارة علمكم لكي يكون البحث خير ممثل لجهد صاحبه،وكلي ثقة في أن ملحوظاتكم ستكون معينة لهذا البحث كي يتخلص من شوائب تعثرت بها يد الباحث،واعذروه في ذلك أنه يحاول أن يجتهد ،آملاً الحصول على اجر المجتهدين،وأعلم أن ملحوظاتكم ما هي إلا نسمات خير تعطر هذه الأطروحة كي تزيدها دقة وإبداعا .

واشكر كل قريب وصديق كان له فضل العون والمساعدة وأدعو الله أن يجزيهم عني خير الجزاء.

فهرس المحتويات

الفصل الخامس:مناقشة لنتائج

المراجع

قائمة الجداول

قائمة الأشكال

قائمة الملاحق

مقارنة إحصائيات فحص جودة مطابقة الفقرة تحت ظروف تباين نوع النموذج الرياضي و طول الاختبار و حجم العينة و التفاعلات بينها

إعداد

راجي عوض الصرايرة

المشرف

الأستاذ الدكتور

خليل عليان

ملخص

هدفت الدراسة للمقارنة بين طرق فحص جودة مطابقة الفقرة ،من خلال استقصاء اثر تباين خصائص الفقرات و نوع النموذج الرياضي و طول الاختبار و حجم العينة و التفاعلات بينها، على معدلات الخطأ من النوع الأول وقوة الاختبار لإحصائيات جودة مطابقة الفقرة ، ولتحقيق هدف الدراسة اعتمد أسلوب المحاكاة في توليد استجابات ثنائية على الفقرات، لأحجام عينات مختلفة من المفحوصين (١٠٠،٥٠٠،١٠٠٠)مفحوص ،وثلاث أحجام مختلفة من طول الاختبار (١٠،٣٠،٦٠) فقرة ،وقد أجريت التحليلات الخاصة للاستجابات لكل عينة من عينات المحاكاة المولدة،ولفحص معدلات الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة تم معايرة (Calibration) البيانات وحساب إحصائي جودة مطابقة الفقرة بنفس النموذج الرياضي الذي استخدم في توليدها، ولفحص قوة الاختبار لإحصائيات جودة مطابقة الفقرة تم معايرة البيانات وحساب إحصائي جودة مطابقة الفقرة بنموذج رياضي يختلف عن الذي استخدم في توليدها .وقد اعتبرت الفقرات التي تجاوزت دلالاتها الإحصائية حدود المطابقة المحددة في الدراسة فقرات غير مطابقة.

وبينت النتائج ما يلي:

- تقل معدلات الخطأ من النوع الأول لإحصائي جودة مطابقة الفقرة G^2 ، χ^2_B بزيادة طول الاختبار ونقصان حجم العينة،في حين تقل معدلات الخطأ من النوع الأول لإحصائي البواقي المعيارية (RMS) بزيادة حجم العينة ونقصان طول الاختبار.
- كشفت النتائج أن قوة الاختبار لإحصائيات جودة مطابقة الفقرة G^2 ، χ^2_B تزداد بزيادة حجم العينة ونقصان طول الاختبار عند جميع مستويات الدلالة ، في حين أن إحصائي البواقي المعيارية (RMS) أبدى معدلات قوة مرتفعة عند زيادة حجم العينة وزيادة طول الاختبار.
- كشفت النتائج أن معدلات الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة تحت افتراض النموذج ثلاثي المعلم أعلى بقليل من معدلات الخطأ من النوع الأول تحت افتراض النموذج ثنائي المعالم بشكل عام وفي معظم الظروف.

الفصل الأول

المقدمة وأهداف الدراسة

الفصل الأول

المقدمة وأهداف الدراسة

سادت النظرية الكلاسيكية في القياس النفسي (Classical test theory) في الفكـر السيكولوجي فترة طويلة، فقد استخدمت أسس هذه النظرية فـي بنـاء أنـواع مختلفـة مـن الاختبارات النفسية والتربوية، وكذلك تحليل البيانات المستمدة من هذه الاختبارات. وبـالرغم من انتشار تطبيق هذه النظرية وما ارتبط بها من مقاييس إحصائية خاصـة بتحليـل فقـرات الاختبار، إلا أن الباحثين في النظرية الحديثة أشاروا إلى جوانب قصور في نظرية الاختبـار الكلاسيكية، التي تحد من استخدامها في الاختبارات الهامة وتطبيقات القياس، ولقد أشار كـل من هامبلتون وسوامنثان (Hambleton and Swaminathan, 1985) إلى عدد من جوانب القصور في النظرية الكلاسيكية ومنها:

١. تعتمد النظرية الكلاسيكية في القياس على استخراج إحصائيات للفقـرة تتمثـل فـي (معامل الصعوبة ومعامل التمييز) وهذا يعتمد على عينة الدراسة المستخدمة وتقتصر على مجتمع الدراسة الذي سحبت منه العينة.
٢. تعتمد درجات الأفراد في اختبار ما على خصائص عينة الفقرات التي يتكـون منهـا الاختبار.
٣. تفترض نظرية الاختبار الكلاسيكية تساوي قيمة الخطأ في القياس لجميع الأفراد الذي يطبق عليهم الاختبار.
٤. خلوها من نماذج رياضية تسهم في تقدير احتمالية الإجابة الصحيحة للفرد علـى أي فقرة من فقرات الاختبار.

ولعل القصور الرئيسي في نظرية الاختبار الكلاسيكية هو عدم اسـتقلالية خـصائص الفقرات عن خصائص المفحوصين أو العكس. وهذا يعني أن خصائص (معالم) الفقرة تعتمـد على عينة المفحوصين الذين استخدموا في تقديرها، وكذلك تعتمد خصائص المفحوصين (معلم القدرة) على عينة الفقرات التي استخدمت في تقديرها.

وجاء استخدام النظرية الحديثة للقيـاس (Item response theory) للتغلـب علـى محددات نظرية الاختبار الكلاسيكية،حيث قدمت طريقة في انتقاء الفقرات تتميز بتقديم معـالم ثابتة للفقرة هي:معالم الصعوبة والتمييز والتخمين،إضافة إلى توحيد المقياس لكل من الصعوبة و قدرة المفحوص لكي يتمكن مطور الاختبار من انتقاء الفقرات الأكثـر ملائمـة لتـصنيف

الأفراد على مستويات الأداء، والتي تتميز في حالة مطابقة البيانات للنموذج، بعدم تغير معالم الفقرات بتغير عينة المفحوصين، وعدم تغير تقديرات القدرات بعينة الفقرات، وبالإضافة لذلك فإنه يمكن مقارنة معالم الفقرات والقدرة لوقوعهما على نفس السلم. فخاصـية اللاتغيـر فـي نظرية استجابة الفقرة تجعل هنالك إمكانية لحل مشكلات القياس والاختبار والتي من الصعوبة حلها في نظرية الاختبار الكلاسيكية، وبشكل خاص: معادلة الاختبار وبنوك الأسـئلة وتحيـز الفقـــــــــــرات والاختبـــــــــــــارات التكيفيـــــــــــــة (Hambleton and Swaminathan and Rogers, 1991).

ففي نظرية استجابة الفقرة يتم الربط بين خصائص الفقرات والقدرة المقاسة من جهـة وبين احتمال الإجابة الصحيحة على الفقرة من جهة أخرى من خلال عـدة نمـاذج رياضـية ومنها النماذج اللوجستيه والتي تستخدم بشكل كبير في نظرية استجابة الفقرة، وهي النمـوذج اللوجستي أحادي المعلم والذي يستخدم فقط معلم واحد لوصف فقرة الاختبار وبشكل خـاص معلم الصعوبة، والنموذج اللوجستي ثنائي المعلـم (Two-Parameter Logistic Model) يستخدم معلمين هما الصعوبة والتمييز، بينما النموذج اللوجـستي ثلاثـي المعـالم يـستخدم الصعوبة والتمييز والتخمين لوصف فقرة الاختبار،ومهما كـان شـكل النمـوذج الرياضـي المستخدم فإن الفائدة منـه تعتمـد بالدرجـة الأولـى علـى مطابقـة البيانـات للنمـوذج . (Lord, 1990; Hambleton and Swaminathan, 1985).

إذ يعتمد نجاح تطبيقات نظرية استجابة الفقرة في تحليل وتفسير نتائج الاختبار علـى أمرين هما تحقق افتراضات النظرية في البيان[ات ومطابقة البيانات للنمـوذج. وإن فحـص مطابقة النموذج قد يسفر عن عدم مطابقة بعض الفقرات للنموذج ،حيث يـشير الأدب علـى أهمية مطابقة البيانات للنموذج (model-data fit)عندما يتم اسـتخدام نمـاذج الاسـتجابة للفقرة. (Dodeen, 2004).

وبشكل عام فإن عدم مطابقة البيانات للنموذج تكون ناتجة عن مجموعة من العوامـل كما وضحها هامبلتون وميكينلي وميلـز (Hambleton, 1993; Mckinely and Mills, 1985).

١. عدم تحقيق افتراضات النموذج المستخدم.

٢. عدم تحقيق خاصية اللاتغير في تقديرات معالم الفقرة والقدرة.

٣. صغر حجم العينة.

وفي هذا الخصوص أشار بجهي Baghi(1990) أن عدم المطابقة يحدث لعدة أسباب، فالفقرات السهلة مثلاً لا تتطابق مع النموذج إذا لم يجب المفحوصين ذوي القدرة المنخفـضة عليها. ويحدث مثل هذا الأمر عادة في اختبارات السرعة خصوصاً عندما لا يكـون هنالـك عقاب على البطئ في الإجابة على فقرات الاختبار. وسبب آخر لعدم المطابقة هـو التخمـين وخاصة في اختبارات الاختيار من متعدد، وهذا النوع من عدم المطابقة يحدث عادةً وبـشكل شائع عندما يجيب المفحوصين من ذوي القدرة المنخفضة على الفقرات الصعبة بشكل صحيح بالتخمين، كما يحدث عدم المطابقة عندما لا يتحقق افتراض أو أكثر من افتراضـات نظريـة استجابة الفقرة، فعلى سبيل المثال إذا كانت فقرات الاختبار تقيس قـدرة أو قـدرات أخـرى إضافة للقدرة المراد قياسها، فهذا انتهاك واضح لافتراض أحادية البعد.

ومن أجل استخدام نماذج نظرية استجابة الفقرة لا بد من فحص كـل مـصدر مـن مصادر عدم المطابقة السابقة (Misfit)، لنتمكن من استخدام فحص جودة المطابقة كمؤشـر على فحص فرضية أحادية البعد، وكذلك الكشف عن الفقرات التي صححت بمفتاح تـصحيح خاطئ (Reise, 1990).

وأشار سميث Smith(1991) أن إحصائي مطابقة الفقرة يزودنا بإطار مرجعي للحكم على الأداء المقدم على الفقرة في الاختبار، أو أداء المفحوص في الاختبار وهذا التحليل يستفاد منه للكشف عن المشاكل التي تتعلق بالقياس مثل التخمين، والأداء التفاضـلي للفقـرة. فعـدم مطابقة الفقرة يدل على أن معالم الفقرة المقدرة تعجز عن اشتقاق نمط إجابـات المفحوصـين عليها مما يجعل تقديرات القدرة التي تم التوصل إليها موضع شك وهذا يعنـي أن القـرارات المبنية على هذه التقديرات غير سليمة. وفي ضوء ذلك فإن فحص جـودة مطابقـة البيانـات لنموذج نظرية استجابة الفقرة ضروري من أجل استخدام النتائج التـي يـتم التوصـل إليهـا باستخدام IRT.

وفي ضوء أهمية جودة مطابقة الفقرة للنموذج فقد تمحورت الدراسـة الحاليـة حـول هـذا المفهوم،وقد برزت الحاجة لتناول عدد من إحصائيات فحص جودة مطابقة الفقرة للنمـوذج و المتمثلة بثلاثة إحصائيات هي:إحصائي بك (Bock) (χ^2_B) ، وإحـصائي (G^2) للأرجحيـة العظمى، وإحصائي البواقي المعيارية(RMS) (Root-Mean Square of The Posterior Deviates)، من اجل المقارنة بين هذه الطرق لتوصل إلى أكثرها فاعلية مـن حيـث قـوة

الاختبار ومعدلات مستوى الخطأ من النوع الأول تحت ظروف تباين نوع النموذج الرياضـي المستخدم وطول الاختبار وحجم عينة المفحوصين وتفاعل هذه المتغيرات بعضها ببعض.

مشكلة الدراسة وأسئلتها:

في ضوء ما سبق يمكن القول أن إحصائيات فحص جودة مطابقة الفقرة للنموذج هي:

إحصائي Bocks (χ^2_B)، وإحصائي (G^2) للأرجحية العظمـى وإحـصائي البـواقي المعيارية (RMS)،. وجاءت مشكلة الدراسة للمقارنة بين هذه الطرق للتوصل إلى أكثرهـا فاعلية من حيث قوة الاختبار ومستوى الخطأ من النوع الأول تحـت ظـروف تبـاين نـوع النموذج الرياضي المستخدم وطول الاختبار وحجم عينة المفحوصين وتفاعل هذه المتغيـرات بعضها ببعض.

وبشكل أكثر تحديداً فإن هذه الدراسة ستجيب عن الأسئلة التالية:

1. هل يختلف الخطأ من النوع الأول وقوة الاختبار لإحصائيات فحص جودة مطابقة الفقرة (إحصائي بك (χ^2_B)، وإحصائي (G^2) للأرجحية العظمى وإحصائي البواقي المعياريـة (RMS))، باختلاف طول الاختبار ؟
2. هل يختلف الخطأ من النوع الأول و قوة الاختبار لإحصائيات فحص جـودة مطابقـة الفقرة (إحصائي بك (χ^2_B)، وإحصائي (G^2) للأرجحية العظمى وإحـصائي البـواقي المعيارية (RMS))، باختلاف حجم العينة ؟
3. هل يختلف الخطأ من النوع الأول لإحصائيات فحص جودة مطابقة الفقرة (إحصائي بك (χ^2_B)، وإحصائي (G^2) للأرجحية العظمى وإحصائي البـواقي المعياريـة (RMS))، باختلاف النموذج الرياضي المستخدم ؟
4. هل يختلف الخطأ من النوع الأول لإحصائيات فحص جودة مطابقة الفقرة (إحصائي بك (χ^2_B)، وإحصائي (G^2) للأرجحية العظمى وإحصائي البـواقي المعياريـة (RMS))، باختلاف التفاعل بين متغيرات: طول الاختبار وحجم العينة؟
5. هل تختلف قوة الاختبار الإحصائي لإحصائيات فحص جودة مطابقة الفقرة (إحـصائي بك (χ^2_B)، وإحصائي (G^2) للأرجحية العظمى وإحصائي البواقي المعيارية (RMS))، باختلاف التفاعل بين متغيرات: طول الاختبار وحجم العينة ؟

التعريفات الإجرائية:

الخطأ من النوع الأول: هو نسبة عدد الفقرات التي تساوت أو قلت قيمة دلالتها عن مستوى الدلالة المحدد واعتبرت فقرات غير مطابقة للنموذج من عدد الفقرات الكلية في مرات إعادة توليد البيانات، والفرضية الصفرية في هذه الحالة صحيحة.

قوة الاختبار: هو نسبة عدد الفقرات التي تساوت أو قلت قيمة دلالتها عن مستوى الدلالة المحدد واعتبرت فقرات غير مطابقة للنموذج من عدد الفقرات الكلية في مرات إعادة توليد البيانات، والفرضية الصفرية في هذه الحالة خاطئة.

أهمية الدراسة:

تتضح أهمية هذه الدراسة في جانبين: الأهمية النظرية وتتلخص في التحقق من النتائج المحتملة بالإجابة عن الأسئلة التي طرحتها الدراسة، أي كيف يمكن استخدام إحصائيات مطابقة الفقرة لتقييم جودة مطابقة الفقرة من اجل تحديد الفقرات غير المطابقة، وفعالية كل إحصائي تحت عدة ظروف هي مستويات حجم العينة وطول الاختبار ونوع النموذج الرياضي. أما من الناحية العملية. فيمكن أن توفر الدراسة معلومات وإرشادات تساعد المهتمين في بناء الاختبارات لانتقاء أفضل الطرق للكشف عن الفقرات غير المطابقة من أجل تعديلها. وإعداد اختبارات متحررة من الفقرات غير المطابقة،وتأتي هذه الدراسة كمصدر لدعم الجهود التي بُذلت لتحديد أفضل الطرق للكشف عن مطابقة الفقرة.

الفصل الثاني

الدراسات السابقة

الفصل الثاني

الإطار النظري و الدراسات السابقة

تناول هذا الفصل جانبين أساسيين في الدراسة، هما الإطار النظري والدراسات السابقة، إذ يتناول الإطار النظري النماذج الرئيسية في نظرية استجابة الفقرة، والتطور التاريخي لتوليد البيانات باستخدام الحاسب الآلي، وكيفية التحقق من مطابقة بيانات القياس لنماذج نظرية استجابة الفقرة، وبشكل خاص التعرف على إحصائيات جودة مطابقة الفقرة ومناقشة الصيغ الرياضية لها وكيفية حسابها. والجانب الثاني يتناول مراجعة الدراسات السابقة ذات الصلة بالأبعاد التي تهتم بها الدراسة، وتشمل دراسات تهدف للمقارنة بين إحصائيات جودة مطابقة الفقرة، والتوصل إلى أكثرها فاعلية من حيث قوة الاختبار أو معدلات الخطأ من النوع الأول أو كلاهما تحت ظروف مختلفة.

الإطار النظري:

نماذج نظرية استجابة الفقرة أحادية البعد:

تفترض النظرية الحديثة في القياس أنه يمكن التنبؤ بأداء الأفراد في اختبار نفسي أو تربوي في ضوء سمة أو خاصية مميزة لهذا الأداء. إذ انبثقت عن هذه النظرية مجموعة من النماذج تعرف باسم نماذج السمات الكامنة (Latent Trait Models) ولكل نموذج معادلة رياضية يتحدد فيها العلاقة بين أداء الفرد على الفقرة والقدرة التي تكمن وراء هذا الأداء وتفسره.

ويمكن التوصل لعدد لانهائي من النماذج اللوجستية والطبيعية، إلا أن النماذج التي حازت على كثير من الاهتمام من جانب الباحثين وعلماء القياس استخدام النماذج اللوجستية بسبب بساطة الصور الرياضية لها مقارنة بالنماذج الأخرى. وترتكن كل من هذه النماذج على عدد من الافتراضات التي يجب أن تتوفر في بيانات الاختبار. وتختلف هذه النماذج في عدد الافتراضات الواجب تحقيقها. ويعتبر النموذج أحادي المعلم أقلها في عدد هذه الافتراضات، ومن هذه الافتراضات التي ترتكن إليها نماذج نظرية استجابة الفقرة. (علام، ١٩٨٦)

أولاً: أحادية البعد (unidimensionality)

ويتمثل هذا الافتراض بوجود قدرة واحدة تفسر أداء الفرد على الفقرة.

ثانياً: الاستقلال المحلي (Local Independence)

وتعني بأن استجابات الفرد على فقرات الاختبار مستقلة إحصائياً عند مستوى قدرة معين، وهذا الافتراض حتى يتحقق يجب أن لا تؤثر إجابة الفرد على فقرة مـا على إجابته على الفقرات الأخرى سواء إيجاباً أو سلباً (Crocker, 1986).

ثالثاً: منحنى خصائص الفقرة (ICC) (Item Characteristic Curve)

ويعرف بأنه اقتران رياضي يحدد العلاقة ما بين احتمال نجاح الفـرد علـى الفقرة وبين القدرة التي يقيسها الاختبار. (Crocker, 1986)

رابعاً: السرعة في الأداء

ويعني أن إجابة الفرد على الفقرة يعود لقدرته، ولا تلعب السرعة أو الوقـت المخصص دوراً في الإجابة على الفقرة (Crocker, 1986).

وبعد مناقشة الافتراضات التي ترتكز عليها نماذج نظرية استجابة الفقرة أحادية البعد، ينبغي توضيح الأسس الرياضية التي بنيت عليها هذه النماذج كمـا بينهـا هـان وهـامبلتون (Han and Hambleton, 2007) وهي:

١. النموذج التجميعي الطبيعي (Normal-Ogive Model).

٢. النموذج اللوجستي أحادي المعلم (1PLM) (One-Parameter Logistic Model)

٣. النموذج اللوجستي ثنائي المعالم (2PLM) (Two-Parameter Logistic Model)

٤. النموذج اللوجستي ثلاثي المعالم (3PLM) (Three-Parameter Logistic Model).

١. النموذج التجميعي الطبيعي (Normal-Ogive Model).

يعد النموذج التجميعي الطبيعي أول نموذج في نظرية استجابة الفقرة. ويعـد نمـوذج ثنائي المعلم، ومنحنى خصائص الفقرة في هذا النموذج اشتقت من اقتران الكثافـة الطبيعيـة (CDF) (Cumulative Density Function) للتوزيع السوي (الطبيعي).

والصورة الرياضية له:

$$p_i(\theta) = \int_{-\infty}^{+a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

حيث:

$P_i(\theta)$: احتمالية إجابة فرد ما تم اختياره عشوائياً من مستوى القدرة θ على الفقرة i إجابة صحيحة.

a_i : معلم التمييز للفقرة i .

b_i : معلم الصعوبة للفقرة i.

z : العلامة المعيارية للفرد متضمنة علامة القدرة ومعلمتي الفقرة.

٢. النموذج اللوجستي أحادي المعلم (One-Parameter Logistic Model) (1PLM)

قدم جورج راش عالم الرياضيات منحى مختلف في نظرية استجابة الفقرة عام ١٩٥٠م. باستخدام الاقتران اللوجستي لاشتقاق منحنى خصائص فقرة (ICC) جديد بدلاً من الاقتران التجميعي الطبيعي (Normal-Ogive Model) الذي يتميز بسهولة الحساب. وقد استخدم هذا النموذج الجديد في تحليل الاستجابة الثنائية، والصيغة الرياضية لحساب احتمالية إجابة فرد ما تم اختياره عشوائياً من مستوى القدرة θ على الفقرة i إجابة صحيحة هي:

$$p_i(\theta) = \frac{1}{1 + e^{-D(\theta - b_i)}}$$

حيث:

e : هي الأساس أللوغاريتمي الطبيعي ويساوي (٢,٧١٨٣).

D : عامل الضبط أو موازنة تدريج القياس (Scaling Factor) وعندما يتم اختيار (D = ١,٧) تقترب قيم $p_i(\theta)$ المقدرة طبقاً للنموذج التجميعي الطبيعي، من القيم المقدرة طبقاً للنموذج اللوجستي ثنائي المعلم (2PL).

ويلاحظ أن المعادلة تتضمن فقط معلم الصعوبة ويفترض أن معلم التخمين يساوي صفر وتساوي التمييز للفقرات.

٣. النموذج اللوجستي ثنائي المعالم (Two-Parameter Logistic Model) (2PLM)

يفترض هذا النموذج أن منحنى خصائص الفقرة يمكن أن يأخذ شكل النموذج اللوجستي التالي.

$$p_i(\theta) = \frac{1}{1 + e^{-Da(\theta - b_i)}}$$

حيث أن a_i : معلمة التمييز للفقرة i ويفترض هذا النموذج أن الأداء علـى الفقـرة يتحدد بناء على معلمتي الصعوبة (b_i) والتمييز (a_i) ويفترض أن معلمـة التخمـين تـساوي صفر.

٥. النموذج اللوجستي ثلاثي المعـالم (3PLM) (Three-Parameter Logistic Model).

يعد هذا النموذج امتداداً للنموذج اللوجستي ثنائي المعالم ولكنه يضيف معلم التخمين إذ يكون الخط التقاربي السفلي لمنحنى خصائص الفقـرة (Lower asymptotes) لا يـساوي صفر، ويرمز له بالرمز (c_i) ويمثل احتمال إجابة الأفراد ذوي القدرة المنخفضة على الفقرة i إجابة صحيحة.

$$p_i(\theta) = C_i + (1 - C_i)\frac{1}{1 + e^{-Da(\theta - b_i)}}$$

مطابقة البيانات للنموذج(Model Data Fit):

يعتبر إجراء مطابقة البيانات للنموذج من الإجراءات الأساسـية فـي نظريـة استجابة الفقرة ،إذ إن فحص جودة مطابقة البيانات للنموذج الرياضي كانت وما زالت موضع اهتمام كثير من الباحثين ،ولقد اقترحت ثلاث طرق رئيسة للتأكد من فحص الجودة كما صنفها هاملتون (Hambleton, 1993) هي:

١. تحديد فيما إذا كانت بيانات الاختبار تطابق افتراضات النمـوذج والمتمثلـة بملائمـة افتراضات النموذج للبيانات.

٢. فحص خصائص النموذج المتوقعة في بيانات الاستجابة للفقرات والمتمثلـة بالفوائـد المتوقعة من استخدام النموذج.

٣. تحديد مدى التقارب والتطابق بين النتائج المتوقعة نظرياً والملاحظة عملياً والمتمثلـة بدقة تنبؤات النموذج. وبالنظر إلى الطرق الثلاث السابقة للمطابقة، يمكن التحقق منها كمـا يلي من خلال (Hambleton and Swaminathan, 1985):

١. تحديد فيما إذا كانت بيانات الاختبار تطابق افتراضات النمـوذج والمتمثلـة بملائمـة افتراضات النموذج للبيانات من خلال.

١. أحادية البعد:

أ. تمثيل بياني للجذور الكامنة في مصفوفة معاملات الارتباط بين الفقرات، والتحقق من توفر عامل عام بحيث يكون الفرق بين الجذر الكامن للعامل الأول والثاني كبيراً.

ب. المقارنة بين رسمين بيانيين لقيم الجذور الكامنة الأول المستخلص بالطريقة السابقة، والثاني المستخلص من معاملات الارتباط بين الفقرات عند تطبيقها في عينة عشوائية.

ج. تحليل البواقي بعد مطابقة نموذج العامل الواحد لمصفوفة التباين المشترك بين الفقرات.

٢. تساوي قيم التمييز في النموذج أحادي المعلم من خلال:

تحليل مدى ارتباط الفقرات بالعلامة الكلية للاختبار، فإذ كان توزيع ارتباطات الفقرة بالعلامة الكلية الاختيار متجانس نختار النموذج الذي يفترض تساوي التمييز.

٣. التخمين:

تدني قيمة هذا المعلم واقترابه من الصفر في النموذجين الأحادي والثنائي المعلم، ويتم ذلك من خلال:

أ. فحص الرسم البياني بين علامة الفقرة وعلامة الاختبار.

ب. الأخذ بعين الاعتبار أداء المفحوصين ذوي القدرات المنخفضة.

٤. الاختبار ليس اختبار سرعة ويتم التحقق منه من خلال:

أ. المقارنة بين تباين عدد الفقرات التي حذفت وتباين عدد الفقرات التي أجيب عليها بشكل غير صحيح.

ب. مقارنة علامات الاختبار عندما يطبق بوقت محدد وعلاماته عندما لا يكون هنالك تحديد للزمن.

ج. فحص نسبة عدد المفحوصين الذين أنهوا الإجابة عن فقرات الاختبار كله، ونسبة عدد المفحوصين الذين أنهوا الإجابة على ٧٥% من الفقرات ونسبة عدد الفقرات التي أجاب عليها نسبة ٨٠% من المفحوصين إذا كانت متقاربة فهذا مؤشر على أن السرعة ليست عاملا مؤثراً.

٢. فحص مدى التقارب بين النتائج المتوقعة والنتائج الملاحظة ويتم من خلال (Hambleton, 1985):

أ. فحص البواقي المعيارية من تحليل بيانات الاختبار (SR_s).

ب. مقارنة العلامات المشاهدة والعلامات المتنبأ بها حسب النموذج، واستخدام الإحصائي مربع كاي للتحقق من دلالة الفروق.

*** فحص خصائص النموذج المتوقعة من خلال (** Hambleton and Swaminathan and Rogers, 1991)**:**

أ. فحص ثبات معالم الفقرات وذلك من خلال مقارنة تقديرات معالم عينة من الفقرات في مجموعتين مختلفين من الأفراد.

ب. فحص ثبات معالم المفحوصين (القدرة) ويتم ذلك من خلال مقارنة تقديرات معالم القدرة لعينة من المفحوصين من خلال تطبيق عينتين أو أكثر من الفقرات المختارة من تجمع من الفقرات

وجاءت طرق النظرية الحديثة في الكشف عن مطابقة الفقرة لمعالجة القصور وعدم الموضوعية الناتج عن استخدام بعض الطرق السابقة ،ولمزيد من التوضيح نجد انه في النظرية الحديثة في القياس كما وضحها Hansen(2004) ،بأن هنالك عدد من الإحصائيات لفحص جودة مطابقة الفقرة (item fit) بعضها يعتمد على إحصائي بيرسون (χ^2) وصيغته الرياضية هي:

$$X^2 = \sum_{j=1}^{J} \frac{(O_j - E_j)^2}{E_j}$$

حيث:

J : عدد الفئات الفرعية للقدرة θ.

j : رقم الفئة الفرعية.

O_{ij} : التكرار الملاحظ للاستجابات الصحيحة في الفئة j.

E_j : التكرار المتوقع للاستجابات الصحيحة في الفئة j.

ويمكن إعادة صياغة المعادلة السابقة بدلالة النسب بدلاً من التكرارات ويصبح إحصائي بيرسون χ^2 :

$$X^2 = \sum_{j=1}^{J} \frac{N_j (O_j - E_j)^2}{E_j (1 - E_j)}$$

حيث:

J : عدد الفئات الفرعية للقدرة θ.

j : رقم الفئة الفرعية.

N_j : عدد الأفراد المستجيبين في الفئة j.

O_j : النسبة الملاحظة للاستجابات الصحيحة في الفئة j.

E_j : النسبة المتوقعة للاستجابات الصحيحة في الفئة j.

وبالمثل فإحصائيات جودة مطابقة الفقرة التي اعتمدت على الأرجحية العظمى في بناء صياغتها الرياضية تكون الصيغة الرياضية لها كالتالي:

$$G^{2} = 2\sum_{j=1}^{J}\left[(O_{j})\ln\left(\frac{O_{j}}{E_{j}}\right) + (1-O_{j})\ln\frac{(1-O_{j})}{(1-E_{j})}\right]$$

حيث:

J : عدد الفئات الفرعية للقدرة θ.

j : رقم الفئة الفرعية.

ln : الاقتران ألوغاريتمي الطبيعي.

O_j : النسبة الملاحظة للاستجابات الصحيحة في الفئة j.

E_j : النسبة المتوقعة للاستجابات الصحيحة في الفئة j. وبشكل عام أساليب إحصائيات فحص جودة مطابقة الفقرة، تتضمن المقارنة ما بين توزيعات العلامة الملاحظة وتوزيعات العلامة المتوقعة عند كل الفئات الفرعية للقدرة، وعملية المقارنة تتضمن عدد من الخطوات:

١. ترتيب المفحوصين حسب تقديرات القدرة تنازلياً أو تصاعدياً.

٢. تصنيف المفحوصين إلى عدد من المجموعات الفرعية بمدى معين.

٣. حساب التكرارات أو النسب الملاحظة للاستجابات الصحيحة عند كل فئة فرعية للقدرة.

٤. حساب النسبة المتوقعة للاستجابات الصحيحة عند كل فئة فرعية للقدرة باستخدام وسيط أو وسط القدرات الواقعة في تلك الفئة.

٥. المقارنة بين توزيعات العلامات الملاحظة والمتوقعة من خلال حساب إحصائي مربع كاي (بيرسون أو الأرجحية العظمى)، وتفحص الدلالة الإحصائية باستخدام توزيع مربع كاي بدرجات حرية مناسبة لكل فقرة لوحدها. ويتم تلخيص التوزيعات الملاحظة والمتوقعة في جدول مطابقة جودة الفقرة للفقرات المصححة ثنائياً (Dichotomous) كما يبين الجدول(١) التالي: (Stone and Zhange, 2003).

الجدول ١. التكرار الملاحظ و المتوقع للاستجابات على الفقرة عند كل فئة فرعية للقدرة

النسبة المتوقعة	(التكرار الملاحظ) النسبة الملاحظة	Group θ
e_1	f_1	1
e_2	f_2	2
.	.	.
ej	fj	J

حيث f_j و e_j هي التكرارات أو النسب الملاحظة والمتوقعة على التوالي و (θ) group تمثل الفئات الفرعية للقدرة. والفرضية الصفرية التي تفحص تنص بأنه "لا يوجد فروق ما بين التكرارات (النسب) الملاحظة والمتوقعة عند كل خلية من خلايا جدول مطابقة جودة الفقرة" هي:

$$H_o : \prod_{jk} = P_k(\theta_j)$$

حيث:

Π_{jk} : تشير لنسبة المستجيبين الملاحظة عند الفئة الفرعية j

$P_k(\theta_j)$: تشير لنسبة المستجيبين المتوقعة عند الفئة الفرعية j التي تحسب من خلال نموذج استجابة الفقرة المستخدم (Hansen, 2004).

وبالاعتماد على هذه الصيغ النظرية، تم تطوير إحصائي بك (Bock) χ^2_B، بالاعتماد على صيغة بيرسون، الذي يستخدم للكشف عن الفقرات غير المطابقة والمصححة ثنائياً والصيغة الرياضية له هي:

$$X^2_B = \sum_{j=1}^{J} \left[\frac{N_j (O_{ij} - E_{ij})^2}{E_{ij}(1 - E_{ij})} \right]$$

حيث:

J : عدد الفئات الفرعية للقدرة θ.

j : رقم الفئة الفرعية للقدرة θ.

i : رقم الفقرة.

N_j : عدد المفحوصين الواقعين في الفئة الفرعية j.

O_{ij} : النسبة الملاحظة للاستجابات الصحيحة على الفقرة i للفئة الفرعية j.

E_{ij} : النسبة المتوقعة للاستجابات الصحيحة على الفقرة i للفئة الفرعية j.

وحيث أن $E_{ij} = \hat{P}_i(\hat{\theta}_{med})$، و $\hat{\theta}_{med}$ هي وسـيط قـيم قـدرة المفحوصـين الواقعين في الفئة الفرعية j . وبعد ترتيب المفحوصين حسب تقديرات القـدرة فـإن إجـراء الإحصائي χ^2_B يصنف المفحوصين إلى عدد من المجموعات الفرعية للقـدرة، ويـتم تحديـد الاستجابات على الفقرة لكل مجموعة فرعية لتحديد توزيـع العلامـة الملاحظـة (التكـرار الملاحظ)، وبالاعتماد على نموذج نظرية استجابة الفقرة المستخدم يتم حساب التوزيع المتوقع (النسبة المتوقعة) ويستخدم الإحصائي χ^2_B وسيط القدرة لقدرات المفحوصين الواقعين في الفئة الفرعية، لحساب النسبة المتوقعة. ويتم هذا الإجراء لجميع الفئات الفرعية J، وتستخدم معادلة χ^2_B لحساب مربع كاي لبك Bock ويتم فحص الدلالـة الإحـصائية للإحـصائي بمقارنتـه بالتوزيع النظري لمربع كاي χ^2 مع درجات حرية تحدد بعدد الفئات الفرعية للقدرة (n) وعدد معالم الفقرة (m): $(d\,f = n - m)$ (Hansen, 2004).

أما إحصائي الأرجحية العظمى الذي طوره ميكلـي وميلـز (1985) Mckinley and Mills

والذي يعتمد على الصيغة النظرية للأرجحية العظمـى، ويعتمـد علـى متوسـط قـدرات المفحوصين في كل فئة فرعية بدلاً من استخدام الوسيط كما هو الحال في إحصائي بك (χ^2_B) ويستخدم هذا الإحصائي للفقرات المصححة ثنائياً (dichotomous item) وصيغته:

$$G^2 = 2\sum_{J=1}^{J}\left[O_{ij}\ \log_e \frac{O_{ij}}{N_J E_i(\bar{\theta}_J)} + (N_J - O_{ij})\ \log_e \frac{N_J - O_{ij}}{N_J(1 - E_i(\bar{\theta}_J))}\right]$$

حيث:

J : عدد الفئات الفرعية للقدرة.

O_{ij} : التكرار الملاحظ للاستجابات الصحيحة للفقرة i في الفئة الفرعية J.

N_J : عدد المستجيبين في الفئة الفرعية J.

$E(\bar{\theta}_J)$: احتمالية الإجابة الصحيحة على الفقرة i عند $\bar{\theta}_J$.

$\bar{\theta}_J$: متوسط قدرة المستجيبين في الفئة J.

وتم تطوير إحصائي (RMS) بالاعتماد على نظرية بيز (Bayes Theorem) التـي تربط ما بين الاحتمالية المشروطة وغير المشروطة (الهامشية) (marginal)،إذ يتم حـساب

البواقي من خلال إيجاد الفرق مابين الاحتمالات البعدية للإجابـة الـصحيحة و الاحتمـالات المتوقعة للإجابة الصحيحة على الفقرة عند القيم المختارة للقدرة،بالاعتماد على أنماط الإجابـة التي يعطيها المفحوصين،فيتم تقسيم المفحوصين إلى فئات بناء على نمط الإجابة على فقـرات الاختبار الكلي ويتم حساب الاحتمالات البعدية للإجابة الـصحيحة و الاحتمـالات المتوقعـة للإجابة الصحيحة على الفقرة عند كل نمط عبر جميع القـيم المختـارة للقـدرة و الـصيغة الرياضية لإحصائي البواقي المعيارية:

$$RMS(\delta_j) = \left[\sum_{k}^{q} \overline{N}k\,\delta^2_{jk} \Big/ \sum_{k}^{q} \overline{N}k \right]^{1/2}$$

إذ تعتبر الفقرة غير مطابقة إذا تجاوزت حد الملائمـة المحـدد، وتـم فحـص أداء الإحصائي في الدراسة تحت ثلاث حدود للملائمة عند Z تساوي: 1، 2، 3 والفقـرات التـي تتجاوز وتزيد عن هذا الحد تعتبر غير مطابقة وتحسب البواقي المعيارية بإيجاد حدود المعادلة كمايلي: (δ_j) تـشير للبـواقي المعياريـة البعديـة للفقـرة J عنـد الفئـة K مـن القـدرة (Standardized Posterior Residual) وتحسب من خلال الصيغة التالية :

$$\delta_{jk} = \frac{\sum_{l}^{s} W_{lk}\,(X_{ij} - P(X_k))}{\left[\sum_{l}^{s} W_{lk}\,(X_{ij} - P(X_k)) \right]^{1/2}}$$

حيث أن W_{lk} هي:

$$W_{lk} = \frac{r_l\;P(X_l / X_k)}{P(X_l)}$$

و X_{ij} : (صفر أو ١).

و $\overline{N}_k$ = عدد المحاولات عند النقطة X_k على متصل القدرة وصيغته الرياضية:

$$\overline{N}_k = \frac{\sum_{l}^{s} r_l\;P(X_l / X_k)\;.\;A(X_k)}{P(X_l)}$$

r_l : تكرارات الإجابة الصحيحة على الفقرة j في النمط X_l في عينة المستجيبين.

s : عدد أنماط الإجابة الملاحظة التي أعطاها المستجيبين.

$P(X_l/X_k)$: الاحتمالية المشروطة لنمط الإجابة X_l عند نقطة القدرة X_k على متصل القدرة.

$A(X_k)$: الأوزان المصاحبة لاقتران الكثافة الـسوي (normal destiny function) عنـد نقطة القدرة X_k على متصل القدرة (التوزيع القبلي للقدرة).

X_k : قيم نقاط القدرة على متصل القدرة والحد الأقصى لعدد هذه القيم التي تجـزئ متـصل القدرة هو:(٢ × $\sqrt{\text{عدد الفقرات}}$).

$P(X_l)$: الاحتمالية المشروطة لنمط الإجابة X_l للمفحوصين (وهي مجموع التوقعات البعديـة عبر جميع نقاط القدرة X_n). (Tiot, 2003).

$$P(X_l) = \sum_{k=l}^{k} P(X_l / X_n) \ . \ A(X_n)$$

وبشكل عام الإحصائيات المستخدمة في سياق نظرية استجابة الفقرة لفحـص جـودة مطابقة الفقرة يتم حسابها من خلال ترتيب المفحوصين حسب القدرة المقدرة لهم، ومـن ثـم تقسيم المفحوصين إلى عدد من المجموعات الفرعية بمدى معين ومن ثم يتم حساب التوزيعات الملاحظة والمتوقعة اعتماداً على وسيط أو متوسط القدرة للمجموعـة الفرعيـة، والتوزيـع للعلامة الملاحظة يتم تحديدُه من خلال تحديد عدد المفحوصين الذين أجابوا بشكل صحيح على الفقرة في كل مجموعة فرعية للقدرة، وتوزيع العلامة المتوقع يتم تحديده من خلال اسـتخدام نموذج استجابة الفقرة للتنبؤ بعدد المفحوصين عند كل فئة فرعية للقدرة، باستخدام وسـيط أو وسط القدرة لتلك المجموعة الفرعية، وتوزيعات العلامة الملاحظة والمتوقعة الموجـودة فـي جدول مطابقة الفقرة. يتم مقارنته من خلال حساب إحصائي مربع كـاي، وتفحـص دلالتـه باسـتخدام توزيــع مربــع كــاي بــدرجات حريــة مناســبة لكــل فقــرة لوحــدها (Stone and Zhange, 2003).

وبناء على ذلك فإن المفحوصين تم تصنيفهم في مجموعات فرعية بناء على القـدرة المقدرة لهم، ولهذا فإن القدرة المقدرة تلعب دوراً كبيراً في حساب توزيعات العلامة الملاحظة والمتوقعة وبالتالي تؤثر على ناتج قيمة إحصائي مربع كاي (Hansen, 2004).

وبالتالي فإن دقة تقدير القدرة تلعب دوراً مهماً في صدق نتائج هذا الإحصائي للكشف عن مطابقة الفقرة، إن الدقة في تقدير القدرة تتأثر بطول الاختبار فعندما يزيد طول الاختبـار تزداد الدقة في تقدير القدرة والنتائج المترتبة على عدم الدقة في تقدير القدرة تنعكس كأخطـاء في تصنيف المفحوصين في جدول مطابقة الفقرة، وهذه الأخطاء تحدث عنـدما يـتم تعيـين

مفحوص فـي مجموعـة فرعيـة قدرتـه الحقيقـة لا تقـع فـي مـدى تلـك المجموعـة (Ankenmann and Stone, 1992).

وقد حذر كل من مونتي وأندرسون Mote and Anderson(1965) مـن عواقـب أخطاء التصنيف التي تقود إلى زيادة في الخطأ من النـوع الأول وضـعف قـوة الاختبـار لإحصائيات مطابقة الفقرة. ومن إحصائيات جودة مطابقة الفقرة الشائعة في تقييم مطابقة الفقرة والتي اعتمدت على مربع كاي مثل إحـصائي (Bock) (χ^2_B) وإحـصائي الارجحيـة(G^2)، وإحصائي البواقي المعيارية (RMS).

وأشارت الأبحاث إلى أن أخطاء التصنيف المرتبطة بدقة تقديرات القدرة تؤثر علـى إحصائيات مربع كاي، وقد أوصى الباحثين بضرورة الأخذ بعين الاعتبار التقـديرات غيـر الدقيقة للقدرة في تقييم مطابقة الفقرة وخاصة للاختبارات القصيرة منهـا (Dodeen, 2004; Stone and Hansen, 2000; Stone, Ankenmann, Lane and Liue, 1995; Stone, and Mazzeo and Mislevy, 1994).

وبناء على ذلك تم تطوير إحصائيات لفحص جودة مطابقة الفقرة لتتناسب و الظروف المختلفة، ومنها إحصائي ستون (Stone) الذي عدل على إحصائي مربـع كـاي باسـتخدام التوزيعات البعدية للقدرة، لزيادة قوة الاختبار وذلك بأخذه بعين الاعتبار عدم الدقة في تقـدير القدرة، وهذا الإحصائي يتضمن بناء جداول لمطابقة الفقرة بالاعتماد على التوزيعات البعديـة لقدرة المفحوصين في حساب الإحصائي ،و يتميز الإحصائي ستون و الـذي يعتمـد علـى التوقعات البعدية للقدرة بأنه لا يمكن فحص دلالته باستخدام التوزيع النظـري لمربـع كـاي، بدرجات حرية محددة، فمن أجل فحص الدلالة يجب إعادة المعايرة باستخدام معامل المعـايرة لكل من درجات الحرية و قيمة الإحصائي بنـاء علـى معـادلات انحـدار خاصـة لكـل منها(Hansen, 2004).

وقد قام ستون وانكيمان ولاين وليو Stone, Ankenmann, Lane and Liue(1995) باستخدام طريقة مونتي كارلو لتوليد توزيعات معاينة تجريبية لإحصائيات المطابقة للفقرة لكل فقرة وتم تحديد القيم الحرجة التي تساعد في فحص الدلالة من التوزيعات التجريبية. وكـذلك وجدوا أن إحصائي المطابقة ستون يتوزع بتوزيع ينتمي لعائلة توزيعات مربع كاي التي يجب أن تستخدم لفحص دلالة إحصائيات مطابقة الفقرة ، ولكن مع إيجاد معامل المعايرة (Scaling Factor) ودرجات الحرية لكل فقرة. ومعامل المعايرة ودرجات الحرية وجـد البـاحثون أن هنالك إمكانية لتقديرها من خصائص بيانات العينة. وعليه فإن فحص الدلالة الإحصائية يـتم

دون الحاجة لتوليد توزيعات تجريبية، ووجد أن خصائص الفقرة مثل التمييز ودالة معلومات الفقرة ومتوسط التباين البعدي ضرورية لتحديد معاملات المعايرة وتستخدم كمتغيرات مستقلة للتنبؤ بمعامل المعايرة.

أما عن طريقة ثيسن وأورلاندو (Thissen and Orlando) لفحص جودة مطابقة الفقرة، فقد تم تطوير إحصائي يعتمد على مجموع العلامات ، إذ أن الإحصائي مطابقة الفقرة الذي تم وصفه من قبل ثيسن وأورلاندو (Thissen and Orlando) يعتمد على إحصائي مربع كاي الذي يستخدم التكرارات الملاحظة والمتوقعة لكل مجموع علامات خام، وهذا الإحصائي لمطابقة الفقرة اعتمد على تقسيم المفحوصين إلى مجموعات بالاعتماد على العلامات الخام. فالعلامة الخام حسب نموذج راش هي إحصائي كافي لتقدير قدرة المفحوص. ويستخدم للفقرات التي ثنائية الاستجابة (Ankenmann, 1994).

توليد البيانات باستخدام الحاسوب:

التطور في علوم الحاسوب، وما صاحب ذلك من انتشار للبرامج التي تساعد الباحثين في مختلف العلوم من إنجاز الأبحاث والتوصل السريع للنتائج، ومن بين هذه البرامج، برنامج توليد البيانات في علم القياس والتقويم، حيث تطورت برامج التوليد والمتعلقة بنظرية استجابة الفقرة منذ ظهور هذه النظرية في السبعينيات.

ومن أجل تحديد خصائص وفاعلية النماذج الرياضية وغيرها من الصيغ الرياضية التي تضمنتها هذه النظرية، فقد اعتمدت البحوث العلمية على أسلوب التوليد (Simulation) من أجل دراسة تأثير المتغيرات الإحصائية المختلفة على أداء إحصائيات نظرية استجابة الفقرة من نماذج وغيرها.

وكانت دراسة المونت كارلو (Monte Carlo) من أكثر الدراسات شيوعاً، وقد استخدمت في الأبحاث التي اهتمت بنظرية استجابة الفقرة وذلك بهدف فحص فاعلية ودقة النماذج الرياضية التي انبثقت عنها والمتعلقة بتقدير معالم الفقرات وقدرة المفحوصين. وكذلك فحص افتراضات أحادية البعد والاستقلال المحلي وفحص طرق كشف ومطابقة البيانات للنموذج المستخدم، وفحص طرق كشف التحيز ومعادلة درجات الاختبار وغيرها.

وقد تبين مدى فائدة هذه الطرق في التوليد للتوصل إلى معلومات دقيقة حول كل إحصائي ونموذج انبثق عن هذه النظرية، وحل كثير من المشكلات المرتبطة بالقياس والإحصاء، وتتميز أساليب التوليد بقدرتها على معالجة المتغيرات مثل الضبط وقدرتها على دراسة تأثير متغيرات متعددة مع بعضها البعض. وإضافة إلى أنها أقل تكلفة من الدراسات

التي تعتمد على تطبيق الأدوات البحثية على الأفـراد، والتوصـل لنتـائج يمكـن تعميمهـا (Harwell, 1996; Han and Hambleton, 2007).

وتطورت أساليب المونت كارلو على ثلاثة مراحل، ففي المرحلة الأولى كانت طرق شكلية للبحث وغير حقيقية، وفي المرحلة الثانية استخدمت في مجال الأبحاث المتعلقة بالقنبلـة النووية، وكانت البداية القوية لاستخدامها بشكل كبير في البحث العلمي، أما في المرحلة الثالثة وكانت في بداية السبعينات فقد تطور علم الحاسب الآلي بشكل كبير، وأصبح متاحاً للعديد من الباحثين، وعندها أصبحت أساليب المونت كارلو طرقاً عامة للبحث وخاصـة فـي المجـال الإحصائي. (Harwell, 1996)

والخطوات العامة المستخدمة في عملية توليد البيانات كالتالي:

١. تحديد المتغيرات التي تتضمنها الدراسة، مثل حجم العينة وطول الاختبار والتوزيع الـذي تقع تحته معالم الفقرات، وغيرها من المتغيرات المراد دراستها.
٢. توليد البيانات وفقاً لنموذج نظرية استجابة الفقرة المطلوب.
٣. تقدير المعالم مثل: معلمة الصعوبة، والتمييز، والتخمين. ومعلمة القدرة باستخدام إحـدى البرامج الخاصة وذلك بالاعتماد على استجابات الفقرة التي تم توليدها بالتوليد.
٤. تكرار الخطوة السابقة (N) من المرات لكل خلية في التصميم البحثي.
٥. مقارنة النتائج التي تقيس تأثير المتغيرات المراد دراستها.
٦. تحليـل النتـائج الخاصـة بكـل خليـة باسـتخدام الطـرق الوصـفية والاسـتنتاجية (Harwell, 1996).
٧. ويظهر الجدول (١) بعض البرمجيات المستخدمة في توليـد الاسـتجابات علـى الفقـرة باستخدام أساليب التوليد، والذي يقارن بين البرمجيات من حيث عدد المفحوصين وعـدد الفقرات الذي يستطيع البرنامج توليد الاستجابات عليها والنماذج الرياضية التي يستخدمها في توليد البيانات ونوع التوزيع (Harwell, 1996).

الجدول ٢.وصف لخصائص برامج توليد البيانات

	WINIRT	PARDSIM	WINGEN2
Examinees	4000	N/A	100000000
Items	400	N/A	100000000
IRT Models	1.Dichotomous	1.Dichotomous	1.Dichotomous 2.Polytomous 3.Multidimensional 4.Nonparametric
Distribution	1.Uniform 2.Normal	1.Uniform 2.Normal	1.Uniform 2.Normal 3.Lognorma4.Two-parameter Beta

ويعد برنامج (Wingen2) أشمل هذه البرامج في توليد الاستجابات على الفقـرات وفقـاً للنماذج المختلفة (سواء نماذج أحادية المعلم، ثنائية المعالم، ثلاثية المعالم)، أو (ثنائيـة أو متعددة الإجابة) أو أحادية البعد أو متعـددة الأبعـاد. ولمختلـف التوزيعـات المنـتظم (uniform) أو الطبيعي (السوي) وغيرها كما هو مبين بالجدول. وحيث يعتبر من أفضل وأهم برامج استجابة الفقرة في محاكاته للبيانات الخاصة (Harwell, 1996).

وقد استخدمت الدراسة البرمجيات التالية:

برنامج (Wingen2):

يستخدم هذا البرنامج في توليد استجابات ثنائية ومتعددة وكذلك اللامعلمية والمتعـددة الأبعاد. وذلك لعينات من الأفراد يولدها البرنامج عشوائياً كي تحاكي عينات المجتمع الأصلي من حيث التوزيعات والخصائص الإحصائية، وهذا البرنامج يتم من خلاله إنتاج العديـد مـن عينات الاستجابة تصل إلـى (١٠٠,٠٠٠,٠٠٠) مفحـوص و(١٠٠,٠٠٠,٠٠٠) فقـرة ذات الأحجام المختلفة التي لها خصائص العينات المناظرة لها بالمجتمع الأصلي نفسها. وبالتـالي يمكن من خلال هذا البرنامج الحصول على عينات عديدة ذات أحجام كبيرة يصعب الحصول عليهـا مـن المجتمـع الأصـلي ممـا يـوفر الوقـت والجهـد والمـال علـى البـاحثين (Han and Hambleton, 2007).

ولقد أُستخدم هذا البرنامج في هذه الدراسة لتوليد العينات مـن الاسـتجابات الثنائيـة (٠،١) لعينات تحاكي عينات المجتمع الأصلي يتراوح حجمهـا مـن (١٠٠) إلـى (١٠٠٠)

مفحوص تحت افتراض التوزيع الطبيعي واختبارات يتراوح طولها من (١٠) فقـرات إلـى (٦٠) فقرة تحت افتراض التوزيع الطبيعي (السوي) لصعوبة الفقرات والتوزيع المنتظم لمعلمة التمييز والتخمين

ولعل التوليد للعينات تعالج العيوب والمشكلات التي تواجه بعض الباحثين في ضـبط الموقف الاختباري مما يمكنهم الحصول على إجابات دقيقة مقارنة بالأدوات البحثية التي تطبق على الأفراد، ومن العيوب والمشكلات عدم الجدية والتخمين وعـدم إكمـال الإجابـة علـى الاختبار. وقد أجريت العديد من الدراسات لمقارنة نتائج عينات من المجتمع الأصلي وأسفرت النتائج أن هنالك ٩٥% من النتائج كانت متطابقة. وهذا مؤشر على مصداقية النتـائج التـي نحصل عليها في مثل هذا النوع من الدراسات (Harwell, 1996).

برنامج (Bilog):

يوفر هذا البرنامج أوامر يمكن من خلالها استخدام النماذج اللوجستية الثلاثة: النموذج أحادي المعلم والنموذج ثنائي المعالم والنموذج ثلاثي المعالم. كما يعطي هذا البرنامج تقديرات لصعوبة الفقرة وكذلك معالم التمييز والتخمين حسب النموذج المـستخدم، إضـافة لتقـديرات القدرة للمفحوصين، ويعطي مؤشرات لجودة مطابقة الفقرة باستخدام إحصائيين هما إحـصائي بك χ^2_B وإحصائي البواقي المعيارية (RMS) ولكل منهما أوامر خاصة يتم كتابتها من أجـل حسابها في مخرجات البرنامج. وهنالك مستوى للمطابقة الإحصائية (Fit) لتوزيع مربع كـاي بهدف التعرف على الفقرات غير المطابقة للنموذج. وتم استخدام ثلاثـة مـستويات لفحـص مطابقة الفقرة وهي (٠,١ ، ٠,٠٥ ، ٠,٠١) وكذلك بالنـسبة للإحـصائي (RMS) البـواقي المعيارية تم استخدام ثلاثة مستويات لفحص مطابقة الفقرة Z=١ ، Z=٢ ، Z=٣، وأي فقرة عند المستوى المراد فحص المطابقة عنده تتجاوز هذا المستوى تعتبر غير مطابقة ويـستخدم هذا البرنامج الوسيط للقدرة في تقدير قيمة إحصائي جودة مطابقة الفقرة.

برنامج (Bilog-Mg):

يماثل برنامج Bilog ولكنه يعطي مؤشرات لجودة مطابقة الفقرة باستخدام إحـصائي الأرجحية العظمي (G^2)، ويستخدم هذا البرنامج وسط القدرة (Mean θ) فـي تقـدير قيمـة إحصائي جودة مطابقة الفقرة

الدراسات السابقة:

فقد تناولت العديد من الدراسات مطابقة البيانات للنموذج الرياضي وللفـرد وللفقـرة والذي يجب ملاحظته على أية حال، إن عدم مطابقة الفقرة لافتراض النمـوذج الرياضـي لا يعني بالضرورة إن نوعية الفقرة ضعيفة. فتحليل مطابقة الفقرة يزودنا بمؤشر تشخيصي حول الفقرة، لهذا السبب فإن عدم مطابقة الفقرة يجب أن لا يكون حله بشطب هذه الفقـرة بـشكل أوتوماتيكي(Seol, 1999).

وضح رايت وبنكابكيسان (Write and panchapakesan(1969 بقولهم "يجب علينا أن لا نفكر بشطب الفقرات بشكل مباشر، التي تكون قيمة مربع كاي لها كبيرة" وهذا يعني أنه يجب علينا أن نقيم ونحدد التأثيرات المحتملة لكل من معالم التمييز والتخمين لهذه الفقرات التي تبدو ضعيفة وغير مطابقة لنموذج استجابة الفقرة المستخدم.

ففي دراسة دودين (Dodeen(2004 التي تهدف إلى فحص أثر معالم الفقرة (التمييز، الصعوبة، ومستوى التخمين) على إحصائي مطابقة الفقرة، والتي استخدمت فيها طريقة التوليد لبيانات فقرات ثنائية الاستجابة (Dichotomous Data) ومن خلال التوليد تم توليـد تـسع اختبارات باستخدام (١٠٠٠) مفحوص و(٥٠) فقرة، وبثلاث مستويات لتمييز الفقرة وتخمينها وصعوبتها.

تم تقدير مطابقة الفقرة باستخدام إحصائيين هما: إحـصائي نـسبة الأرجحيـة (χ^2_B) والبواقي المعيارية (RMS) ، وأشارت النتائج إلى أن مستويات تمييز الفقرة والتخمـين أثـر على قيم مطابقة الفقرة، إذ أنه بزيادة مستوى تمييز الفقرة أو التخمين أدى إلى زيادة في قـيم إحصائيات مطابقة الفقرة، ووجد أن مستويات صعوبة الفقرة لا تؤثر على إحـصائي مطابقـة الفقرة.

في دراسة روجرز وهيتي (Rogers and Hattie(1987 والتي درسوا فيها ثـلاث إحصائيات لمطابقة الفقرة ومطابقة الفرد لنموذج استجابة الفقرة أحادي المعلم (1PL)، وكانت: متوسط مربع البواقي (Mean- Square residual) ومجموع –ت (Total –t) ومؤشر –ت (between –t index)، إذ تم استخدام بيانات توليد لـ(٥٠٠) مفحوص و(١٥) فقرة. ووجد الباحثون أن إحصائي (between –t index) حساس للتخمين (عـدم التجـانس فـي معلـم التخمين). وأوصى الباحثون باستخدام هذا الإحصائي في فحص مطابقة النموذج.

كذلك أجرى فيرش (Farish(1984 دراسة لأثر اختلاف حجم العينة وطرق اختيارها على مطابقة الفقرة لنموذج راش، كما درسوا أثر إزالة الفقرات غير المطابقة علـى مطابقـة الاختبار للنموذج. استخدم الباحث طرق لاختيار العينة هي (العينة العشوائية البسيطة، ثـلاث عينات عنقودية مع اختلاف العناقيد)، وطبقت الدراسة على أكثر مـن (٢٠٠٠) طالـب فـي استراليا. وأشارت النتائج إلى أنه بزيادة حجم العينة تزداد قوة إحصائي مطابقة الفقرة. وبالمثل وجد أن إزالة الفقرات غير المطابقة تحسن من مطابقة الاختبار بينما بقاء هذه الفقرات تجعـل المطابقة ضعيفة.

أما دراسة هامبلتون وآخرون (Hambleton et al.(1991 والتـي هـدفت لفحـص حساسية إحصاء جودة المطابقة لحجم العينة، إذ استخدمت الدراسـة عينـة بحجـم (٢٤٠٠) مفحوص استجابوا على (٥٠) فقرة، تم توليد استجاباتهم باستخدام نموذج ثنائي المعلم (2PL) وتم استخدام إحصائي مطابقة الفقرة (Q1) مع خمس أحجام من العينات التي تمت محاكاتهـا (١٥٠، ١٤٠، ٤٠٠، ٦٠٠، ٢٤٠٠) وأشارت النتائج إلى أن الزيادة في حجم العينـة يـؤدي لزيادة عدد الفقرات غير المطابقة.

أما دراسة مونت وسكيوميكر (Movnt and Schumacker(1998 والتي تناولـت أثر التخمين على ثلاثة إحصائيات لمطابقة الفقرة وهـم:(weighted total, unweighted total, un weighted between fit) . تم توليد بيانات باستخدام التوليد لـ ـ (١٠٠) فقـرة و(١٠٠) مفحوص. التي تطابق نموذج راش، والذي تم استخدامه في التوليد ثلاثة مـستويات من التخمين (صفر، ٢٥%، ٥٠%) وتوزيعين لصعوبة الفقرة (uniform, normal). أشـار الباحثون بأنه لا يوجد هناك أثر لمستوى التخمين على متوسط إحصائي مطابقة الفقرة لكـل نوع من توزيعات صعوبة الفقرة. وبالمثل إحصائي المطابقة (between fit statistic) كـان أكثر حساسية لاحتمالية التخمين، وعدد الفقرات التي حددت بأنها غيـر مطابقـة باسـتخدام (between fit) زادت مع زيادة مستوى التخمين. وأوصى البـاحثين باسـتخدام إحـصائي (between fit) كإحصائي جيد لتحليل الفقرة وبناء الاختبار.

أما عن دراسة ريسي (Reise(1990 استخدم ثمانية مجموعـات لبيانـات التوليـد، (١٠٠٠) مفحوص و(٥٠) فقرة من نوع الاختيار من متعدد. وتم المقارنة بـين إحـصائيات مطابقة الفقرة (χ^2_B) مع مؤشر مطابقة الفرد (Z_3) لكل فقرة. إذ درسوا أثـر مجموعـة مـن العوامل على أداء كل مؤشر، وتبين من الدراسة أن العلاقة ما بين معلم التمييز وقيم التخمـين

ومطابقة الفقرة لم تحدد ولم تعرف، وبالمثل معام الصعوبة أشارت النتائج أنه لا توجد علاقـة بين صعوبة الفقرة ومطابقة الفقرة.

ودرس سيول (Seol(1999 القوة لخمسة مؤشرات معيارية لمطابقة الفقـرة للكـشف عن الأداء التفاضلي للفقرة (DIF) في نموذج راش. والخمسة مؤشرات هي:

(weighted standardized, un weighted standardized indices)

(UWz , Wz)

(Extended Caution indices), (standardized Likelihood index).

والمؤشرات الخمسة بصورتها الأصلية تستخدم لتقدير مطابقة الفرد في نموذج راش، إذ تم استخدامها لتقدير والكشف عن مطابقة الفقرة، باسـتخدام بيانـات توليـد لـ ـ (١٠٠٠) مفحوص و(١٠٠) فقرة. وأشارت النتائج أنه لا يوجد علاقة بين قيم المؤشرات الـسابقة مـع معام صعوبة الفقرة.

ففي دراسة لتوليد بيانات أجراها أورلاندو وثيسن (Orlando and Thissen(2000 التي هدفت لتقييم أداء إحصائيات مطابقة الفقرة ($S - \chi^2$) و ($S - G^2$) على ثلاثة مـستويات لطول الاختبار (١٠، ٤٠، ٨٠) فقرة، وتم توليد البيانات تحت النماذج الثلاثة لنظرية استجابة الفقرة (نموذج أحادي المعلم، ونموذج ثنائي المعلم، ونموذج ثلاثي المعلم) ولتقييم قوة الاختبار الإحصائي تم توليد البيانات تحت افتراضات نموذج وتدريج استجابات الفقرة بنموذج مختلـف عن النموذج المستخدم في توليد البيانات. وأشـارت النتـائج أن الخطـأ مـن النـوع الأول للإحصائي ($S - \chi^2$) عبر جميع أحجام العينة كانت ثابتة، ولكن الخطـأ مـن النـوع الأول للإحصائي ($S - G^2$) زادت بازدياد حجم العينة و قلت بزيادة طـول الاختبـار. وأظهـرت النتائج أيضا أن قوة الاختبار الإحصائي ($S - \chi^2$) لمطابقة الفقرة عالية نسبيا في الكشف لعدم المطابقة. عند تم توليد البيانات تحت النموذج الثنائي المعلم والنموذج الثلاثي المعلم واستخدام النموذج أحادي المعلم في فحص المطابقة. ولكن قوة الاختبار ($S - \chi^2$) غير مناسبة للكـشف عن عدم المطابقة لبيانات تم توليدها تحت النموذج ثنائي المعلم واستخدام النموذج ثلاثي المعلم في فحص المطابقة، وأشار الباحثون أن استخدام الإحصائي ($S - \chi^2$) واعد في كشف عـدم المطابقة للفقرات المصححة ثنائيا.

ودراسة ستون وزانج (Stone and Zhang (2003 قـارنوا بـين إحـصائي (χ^{2*}) وطريقة اورلاندو وثيسن(Orlando and Thissen) وطريقةدونيو وهومبـو (Donoghue and Hombo) تحت ظروف مختلفة لأحجام العينة وطول الاختبار، وكان معيار المقارنة هو

القوة الإحصائية لاختبارات مطابقة الفقرة والخطأ من النوع الأول المرتبطة بها. حيـث تـم استخدام فقرات مصححة ثنائيا. والفقرات غير المطابقة تم توليدها من خلال توليد بيانات تحت افتراضات نموذج وعمل تدريج لاستجابات الفقرات بنموذج آخر. وأيـضا توليـد اسـتجابات الفقرة بمعالم فقرات، وتغير بعض المعالم لبعض الفقرات عند القيام في المعايرة.

وأشارت النتائج أن إحصائي χ^{2*} وإحصائي (Orlando and Thissen) لهما نفـس الخطأ من النوع الأول بغض النظـر عـن حجـم العينـة وطـول الاختبـار، وإحـصائي (Donoghue and Hombo) لم يطبق على الاختبارات ذات الأطـوال (٢٠، ٤٠) فقـرة. بسبب المحددات العملية لهذا الإجراء ،وتم تطبيقه فقط على الاختبار المكـون مـن عـشرة فقرات، وأشارت النتائج المتعلقة بهذا الإحصائي أن الخطأ من النوع الأول لاختبـار العـشرة فقرات كانت أقل من المستوى الذي حدده الباحثون للخطأ من النوع الأول،وكـذلك أوضـحت نتائج الدراسة أن الخطأ من النوع الأول للإحصائيات التي لا تعتمد على التوزيعات البعديـة تقل بزيادة طول الاختبار ونقصان حجم العينة و زادت قوة الاختبـار بزيـادة حجـم العينـة ونقصان طول الاختبار ،في حين أن إحصائي ستون والذي يعتمد علـى التوقعـات البعديـة لتوزيع القدرة تقل الخطأ من النوع الأول فيه بزيادة حجم العينة وزيادة طول الاختبار فـي حين زادت قوة الاختبار بزيادة حجم العينة ونقصان طول الاختبار،وإضـافة لـذلك أن قـوة الاختبار زادت في حالة استخدام النموذج أحادي المعلم في توليد لبيانات و استخدام النموذج ثلاثي المعلم في تدريج البيانات منها عند استخدام النموذج ثنائي المعلم في توليد لبيانـات و استخدام النموذج ثلاثي المعلم في تدريج البيانات، وتمت مقارنة هـذه الإحـصائيات أيـضا بإحصائي (Bock) الذي أعطى أن كل الفقرات تقريبا غير مطابقة وخاصة في الاختبـارات الصغيرة، بمعنى أن الخطأ من النوع الأول المرتبطة بهذا الإحصائي كبيرة. ومـن الملاحـظ بالإضافة لذلك زيادة الخطأ من النوع الأول مع زيادة حجم العينة، ولكن تقل الخطأ من النوع الأول مع زيادة طول اختبار. وهذا دليل على أن الطرق التقليدية لها معدلات خطأ من النـوع الأول كبيرة، أما بالنسبة لقوة الاختبارات الإحصائية، بشكل عام إحصائي (χ^{2*}) أبـدى قـوة أكثر من الطرق الأخرى. وإحصائي (Orlando and Thissen)تميز بقوة مناسبة لأحجـام العينات من (٢٠٠٠) مفحوص عندما تم استخدام (3P/2P) في توليد البيانات وتـم اسـتخدام (1P) في المعايرة وقوة هذا الإحصائي تعتمد على معلم تمييز الفقرة وبشكل عام زادت قـوة الاختبار مع زيادة حجم العينة ، أما إحصائي (Donoghue and Hombo) الـذي اسـتخدم

فقط لاختبار من طول عشرة فقرات، فعند هذا الطول قوته مشابهة لإحصائي (Orlando and Thissen) وكلاهما قوتهم أقل من قوة الإحصائي (χ^{2*}).

ومن الدراسات التي اعتمدت قوة الاختبار للمقارنة بين إحـصائيات مطابقـة الفقـرة دراسة ين Yen(1981)، إذ قدم دراسة توليد لمقارنة بين الإحصائي (Q_1) وإحصائي الـذي وضع من قبل Write and Panchapakesan(1969)، حيث تم توليد استجابات على (٣٦) فقرة و(١٠٠٠) مفحوص تحت النموذج أحادي المعلم والنموذج ثنائي المعلم والنموذج ثلاثـي المعلم. وجد الباحث تشابه النتائج المستخلصة من الاختبارين. ولاحـظ أيـضا عـدم قـدرة الإحصائي على كشف عدم المطابقة للفقرة عندما تم توليد البيانات تحت النموذج الثلاثي، وتم تدريج الاستجابات باستخدام النموذج الثنائي. وقام بفحص التقارب بين توزيع الإحـصائي Q_1 مع التوزيع النظري لمربع كاي، وقد وجد أن الإحصائي Q_1 توزيعه يتقـارب مـع التوزيـع النظري لمربع كاي.

وأجرى كل من ميكينلي وميلز Mckinely and Mills(1985) مقارنة بين إحصائي Bock's (χ^2_B) وإحصائي Yen (Q_1) والإحصائي LCHI بدراسة توليد، واستخدام الباحثون أحجام العينات (٥٠٠، ١٠٠٠، ٢٠٠٠) مفحوص وثلاثة توزيعات للقدرة (منخفض، متوسط، عالي) وأربعة نماذج لنظرية استجابة الفقرة (النموذج اللوجستي أحـادي المعلـم، والنمـوذج اللوجستي ثنائي المعلم، والنموذج اللوجستي ثلاثي المعلـم، و Two factor linear) وكـان طول الاختبار المستخدم (٧٥) فقرة، وأشارت النتائج إلى أن إحصائي (LCHI) يتمتع بخطـأ من النوع الأول أقل من إحصائي بك وإحصائي Q_1. واقترح الباحثون أن اختيـار إحـصائي المطابقة يجب أن يعتمد على الخطأ من النوع الأول.

وكما أجرى أنسلي وبي Ansley and Bae(1989) دراسة توليد، لفحـص توزيـع المعاينة للإحصائي (Q_1) تحت النموذج ثلاثي المعالم، لفقرات اختبار مصححة ثنائيا. حيث تم استخدام اختبارات بطول (٣٠) فقرة و(٦٠) فقـرة وحجـم عينـة مـن (١٠٠٠ و ٢٠٠٠) مفحوص. ووجد أن الإحصائي (Q_1) لا يتقارب توزعه مع التوزيع النظـري لمربـع كـاي بدرجات حرية (J - m).

وأيضا دراسة ستون وهانسن Stone and Hansen(2000)، التي اهتمت في فحص التوزيع النظري لإحصائيات مطابقة الفقرة لمربع كاي لبيرسون والأرجحيـة العظمـى. (χ^2) و(G^2) على التوالي. تحت نموذج الاستجابة المتدرجة (GRM) لاختبار يتكـون مـن (٣٢)

فقرة، وأشارت النتائج إلى أن توزيع المعاينة لهذه الإحصائيات يقارب توزيـع مربـع كـاي النظري ولكنه ينحرف عن توزيع مربع كاي عندما يكون طول الاختبار قصير (٨) فقرات أو (١٦) فقرة، ومن الملاحظ أن الدراسات السابقة أشارت إلى أن توزيعات المعاينة لإحصائيات مطابقة الفقرة التقليدية يتقارب مع التوزيع النظري لمربع كاي في الاختبارات الطويلـة التـي ترتبط مع التقديرات الدقيقة لمعلم القدرة في حين تكون توزيعاتها أكثر انحرافا عـن التوزيـع النظري لمربع كاي في الاختبارات القصيرة والمرتبطة بالتقديرات غير الدقيقة للقدرة.

ففي دراسة ستون وهانسن (Stone and Hansen(2000 التي هدفت لفحـص أثـر الدقة في تقدير القدرة على توزيع إحصائيات مربع كاي لجـودة المطابقـة. تحـت نمـوذج الاستجابة المتدرج (GRM). وقد قام الباحثون في توليد توزيعات معاينة تجريبية لإحصائيات المطابقة تحت عدة متغيرات: طول الاختبار وحجم العينة. حيث استخدمت الدراسة اختبارات بطول (٨) (١٦) (٣٢) فقرة وعينات بأحجام (١٠٠٠) و(٢٠٠٠) مفحوص، ومتغيـر طـول الاختبار تم استخدامه لتوضيح أثر الدقة في تقدير القدرة على إحصائيات مطابقة الفقرة ومتغير حجم العينة، للكشف عن حساسية اختبار جودة مطابقة الفقرة لحجم العينة. وتم حساب كل من إحصائيات مطابقة الفقرة لمربع كاي لبيرسون والأرجحيـة العظمـى. وتوزيعـات المعاينـة التجريبية لهذه الإحصائيات ولدت من ١٠٠٠ تكـرار (replications) ولتقيـيم توزيعـات إحصائيات جودة المطابقة، استخدم الباحثون الرسومات البيانية (Q-Q) للمقارنـة مـا بـين التوزيعات التجريبية مع التوزيع النظري لمربع كاي. وكذلك للمقارنة ما بين الخطأ من النوع الأول ومقارنة الإحصائيات الوصفية لتوزيعات المعاينة، وقد أشارت النتائج لوجود انحـراف بسيط لتوزيعات التجريبية عن التوزيع النظري لمربع كاي، ففي حالة الاختبار الذي يتكون من (٣٢) فقرة و(١٠٠٠) مفحوص فإن التوزيعات التجريبية لإحصائي مطابقـة الفقـرة مقاربـة ومشابهة لتوزيع مربع كاي النظري، وفي حالة الاختبارات القصيرة (٨)، (١٦) فقرة فهنالـك فروق بين التوزيعات التجريبية لإحصائيات بيرسون عن التوزيع النظري، وهناك زيادة فـي الخطأ من النوع الأول. وبكلمات أخرى عدم الدقة في تقدير القدرة. يولد اختلاف وفروق مـا بين التوزيعات التجريبية والتوزيعات النظرية لإحصائي المطابقة لبيرسون. وعنـدما يكـون هناك تقدير دقيق للقدرة فإن التوزيع التجريبي لإحصائيات مطابقة الفقرة يقترب من التوزيـع النظري. وكذلك الحال بالنسبة لإحصائي المعتمد على الأرجحية أظهر نتـائج بـشكل عـام مشابهة، ولقد أوصى الباحثين أن عدم الدقة في تقدير القدرة يجب أن يؤخذ بعين الاعتبار عند استخدام إحصائيات مطابقة الفقرة التقليدية وخاصة في الاختبارات القصيرة.

أما فيما يتعلق بالدراسات التي اهتمت بالإحصائيات البديلة التي تعتمد على التوزيعات البعدية لفحص جودة مطابقة الفقرة، فقد أجرى ستون وزانـج Stone and Zhang (2003) دراسة توليد لفحص قوة الاختبار الإحصائي والخطأ من النوع الأول لكل من إحصائي (G^{2*}) و (χ^{2*}) واستخدمت الدراسة اختبارات بطول (٦) و (١٢) فقرة، توزيعات للقـدرة (توزيـع طبيعي وتوزيع ملتوي التواء موجبا). وحجم عينـة (٥٠٠، ١٠٠٠، ٢٠٠٠) مفحـوص تـم استخدام معالم الفقرات لاختبـار (National Assessment of Educational) (NAEP Progress)، أشارت الدراسة إلى أن البيانات التي تمت محاكاتها تحـت افتـراض التوزيـع الطبيعي للقدرة. أن الخطأ من النوع الأول قاربت المستويات التي حددها الباحث للاختبـارات التي تتكون من (٦) فقرات و(١٢) فقرة بغض النظر عن حجم العينة، بالإضافة إلى أن الخطأ من النوع الأول لإحصائيات جودة المطابقة (χ^{2*}) (G^{2*}) متشابهة. وعندما تم توليد البيانـات في التوزيع الملتوي للقدرة فإن الخطأ من النوع الأول زادت عن المـستويات التـي حـددها الباحث في كل الظروف. وإن الخطأ تزداد بزيادة حجم العينة ونقصان طول الاختبـار، أمـا النتائج المتعلقة بقوة الاختبار وعند التوزيع الطبيعي للقدرة وجد أن القوة تكون كبيرة عند حجم العينة (١٠٠٠) و(٢٠٠٠) وبشكل عام وجد الباحث أن القوة تقل عنـدما تقـل (α)، وتـزداد عندما يزداد حجم العينة ونقصان طول الاختبار ، وأيضا وجد أن قوة الاختبار G^{2*} أقل بشيء بسيط من قوة الاختبار χ^{2*} ، وأشار الباحث إلى أن الإحصائيين السابقين يمكن أن يـستخدموا للكشف عن الفقرات غير المطابقة للاختبارات التي تتضمن تقديرات غير دقيقة للقدرة.

في ضوء هذا العرض للدراسات السابقة تتضح أن إحصائيات جودة مطابقـة الفقـرة تختلف فيما بينها من حيث قوة الاختبار والخطأ من النوع الأول، تبعا لعدد من المتغيرات التي تتعلق بطول الاختبار وحجم العينة ، ومن خلال الاطلاع على الدراسـات الـسابقة نجـد أن الدراسات السابقة اقتصرت على دراسة متغير أو متغيرين من المتغيرات السابقة على الخطـأ من النوع الأول أو قوة الاختبار. وجاءت هذه الدراسة لفحص تأثير هذه المتغيرات مجتمعـة مع بعضها بالإضافة إلى متغير آخر وهو نوع النموذج الرياضي وتفاعلات هذه المتغيرات مع بعضها البعض على قوة اختبار إحصائي مطابقة الفقرة والخطأ من النوع الأول ، حيث تتـيح المجال لاختبار فاعلية هذه الإحصائيات مجتمعة مع بعضها تحت نفس الظروف لاسـتخلاص إيجابيتها وسلبياتها، إذ تعتبر هذه الدراسة استكمالا للبحوث والدراسات المعاصرة في الكـشف عن فاعلية إحصائيات مطابقة الفقرة التي تعد مجالا خصبا للبحث.

الفصل الثالث

الطريقة والإجراءات

الفصل الثالث

الطريقة والإجراءات

يتناول الفصل وصفاً للطرق والإجراءات المختلفة، التي أتبعت للإجابة عـن أسـئلة الدراسة، ويتضمن عرضاً لأساليب و إحصائيات فحص جودة مطابقة الفقرة وطرق حسابها ، والمتغيرات المؤثرة عليها، وكيفية توليد بيانات الاستجابة على الفقرات،والطريقة التي يتم فيها فحص الخطأ من النوع الأول، وقوة الاختبار لإحصائيات جودة مطابقة الفقرة المستخدمة فـي الدراسة.

أولاً- إحصائيات جودة مطابقة الفقرة:

تناولت هذه الدراسة ثلاثة إحصائيات لفحص جودة مطابقة الفقرة وهي: إحصائي بـك (Bock) χ^2_B، وإحـصائي الأرجحيـة العظمـى (G^2)، وإحـصائي البـواقي المعياريـة (RMS).وفيما يلي عرضا للبرامج المستخدمة في حساب كل إحـصائي مـن الإحـصائيات الثلاثة المذكورة:

- يحسب إحصائي جودة مطابقة الفقرة بك (Bock) (χ^2_B) من خلال معادلة مربـع كاي لبكχ^2_B ،والتي تعتمـد بـشكل عـام علـى المقارنـة مـا بـين توزيعـات العلامـة الملاحظة(التكرارات الملاحظة) وتوزيعات العلامة المتوقعة (التكرارات المتوقعة) في كل فئة من الفئات الفرعية للقدرة، وعملية المقارنة تتضمن عدد من الخطوات التالية:

١. ترتيب المفحوصين حسب تقديرات القدرة تنازلياً أو تصاعدياً.
٢. تصنيف المفحوصين إلى عدد من المجموعات الفرعية بمدى معين.
٣. حساب التكرارات أو النسب الملاحظة عند كل فئة فرعية للقدرة.
٤. حساب النسبة المتوقعة عند كل فئة فرعية للقدرة باستخدام وسيط القدرات الواقعة في تلك الفئة.
٥. المقارنة بين توزيعات العلامات الملاحظة(التكرارات الملاحظة) و توزيعات العلامـات المتوقعة (التكرارات المتوقعة)من خلال معادلة مربع كاي لبكχ^2_B ،ويتم فحـص الدلالـة الإحصائية للإحصائي بمقارنته بالتوزيع النظري لمربع كاي χ^2 مع درجات حرية تحـدد بعدد الفئات الفرعية للقدرة (n) وعدد معالم الفقرة (m): (d f =n-m).

وتم حساب إحصائي جودة مطابقة الفقرة بك (Bock) (χ^2_B) في جميـع الظـروف المختلفة لهذه الدراسة ،من خلال برنامج Bilog ،وذلك بكتابة الأوامر (Syntax) الخاصة به ، والتي تظهر في الملحق (١).

- يحسب إحصائي جودة مطابقة الفقرة (G^2) (إحصائي الأرجحية العظمـى) مـن خلال المعادلة ، والتي تعتمد على صيغة النظرية للأرجحية العظمى،،والتي تعتمد بشكل عـام على المقارنة ما بين توزيعات العلامة الملاحظة(التكرارات الملاحظة) وتوزيعـات العلامـة المتوقعة (التكرارات المتوقعة) في كل فئة من الفئات الفرعية للقدرة، وعملية المقارنة تتضمن عدد من الخطوات التالية:

١. ترتيب المفحوصين حسب تقديرات القدرة تنازلياً أو تصاعدياً.
٢. تصنيف المفحوصين إلى عدد من المجموعات الفرعية بمدى معين.
٣. حساب التكرارات أو النسب الملاحظة عند كل فئة فرعية للقدرة.
٤. حساب النسبة المتوقعة عند كل فئة فرعية للقدرة باستخدام وسط القدرات الواقعة في تلـك الفئة.
٥. المقارنة بين توزيعات العلامات الملاحظة(التكرارات الملاحظة) و توزيعات العلامـات المتوقعة (التكرارات المتوقعة) من خلال معادلته ،والتي تعتمد علـى الـصيغة النظريـة للأرجحية العظمى ،ويتم فحص الدلالة الإحصائية للإحصائي بمقارنته بالتوزيع النظـري لمربع كاي χ^2 مع درجات حرية تحدد بعدد الفئات الفرعية للقدرة (n) وعدد معالم الفقرة (m): (d f =n-m).

وتم حساب إحصائي جودة مطابقة الفقرة (G^2) (إحصائي الأرجحية العظمى) فـي جميع الظروف المختلفة لهذه الدراسة من خلال برنامج Bilog-MG، وذلك بكتابة الأوامـر (Syntax) الخاصة به ، والتي تظهر في الملحق (٢).

يحسب إحصائي (RMS) بالاعتماد على نظرية بيز (Bayes Theorem) التي تربط ما بين الاحتمالية المشروطة وغير المشروطة (الهامشية) (Marginal)،إذ يتم حساب البواقي من خلال إيجاد الفرق مابين الاحتمالات البعدية للإجابة الصحيحة و الاحتمـالات المتوقعـة للإجابة الصحيحة على الفقرة عند القيم المختارة للقدرة بالاعتماد على أنمـاط الإجابـة التـي يعطيها المفحوصين على فقرات الاختبار ،فيتم تقسيم المفحوصين إلى فئات بناء علـى نمـط الإجابة على فقرات الاختبار الكلي

ويتم حساب الاحتمالات البعدية للإجابة الصحيحة و الاحتمـالات المتوقعـة للإجابـة الصحيحة على الفقرة عند كل نمط عبر جميع القيم المختارة للقدرة .

وتم حساب إحصائي جودة مطابقة الفقرة (RMS) البواقي المعيارية من خلال برنامج Bilog في جميع الظروف المختلفة لهذه الدراسة وذلك بكتابة الأوامر (Syntax) الخاصة به والتي تظهر في الملحق (٣).

ثانيا:تصميم الدراسة ومتغيراتها:

أ. <u>متغير طول الاختبار:</u> حدد لهذا المتغير ثلاث مـستويات هـي (١٠، ٣٠، ٦٠) فقـرة، ويوضح لنا المتغير بمستوياته المختلفة أثر طول الاختبار على الخطأ مـن النـوع الأول وقوة الاختبار لإحصائيات جودة مطابقة الفقرة من أجل تحديد الإحصائي المناسب عند كل مستوى يجب استخدامه.

ب. <u>متغير حجم العينة:</u> حدد لهذا المتغير ثلاث مستويات هي (١٠٠، ٥٠٠، ١٠٠٠) مفحوص ويوضح لنا المتغير بمستوياته المختلفة، أثر حجم العينة على الخطأ من النوع الأول وقوة الاختبار لإحصائيات جودة مطابقة الفقرة من أجل تحديد الإحصائي المناسب عنـد كـل مستوى يجب استخدامه.

ج .<u>نوع النموذج الرياضي:</u> حدد لهذا المتغير مستويين هما النموذج اللوجستي ثنائي المعـالم، والنموذج اللوجستي ثلاثـي المعـالم فـي عمليـة توليـد البيانـات وكـذلك فـي عمليـة معايرتها(Calibration) لكل من الفقرات والمفحوصين.

ثالثا- توليد البيانات Simulating the data :

تم توليد بيانات الاستجابة على الفقرات باسـتخدام برنـامج (Wingen2). لتـشمل المتغيرات المختلفة لهذه لدراسة، و ذلك للمقارنة مابين إحصائيات فحص جودة مطابقة الفقرة تبعا لمعيارين هما :الخطأ من النوع الأول و قوة الاختبار وفيما يلي عرضـا لكيفيـة توليـد البيانات.

أ- توليد البيانات المتعلقة بفحص الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة تحت افتراض النموذج اللوجستي ثنائي المعالم.

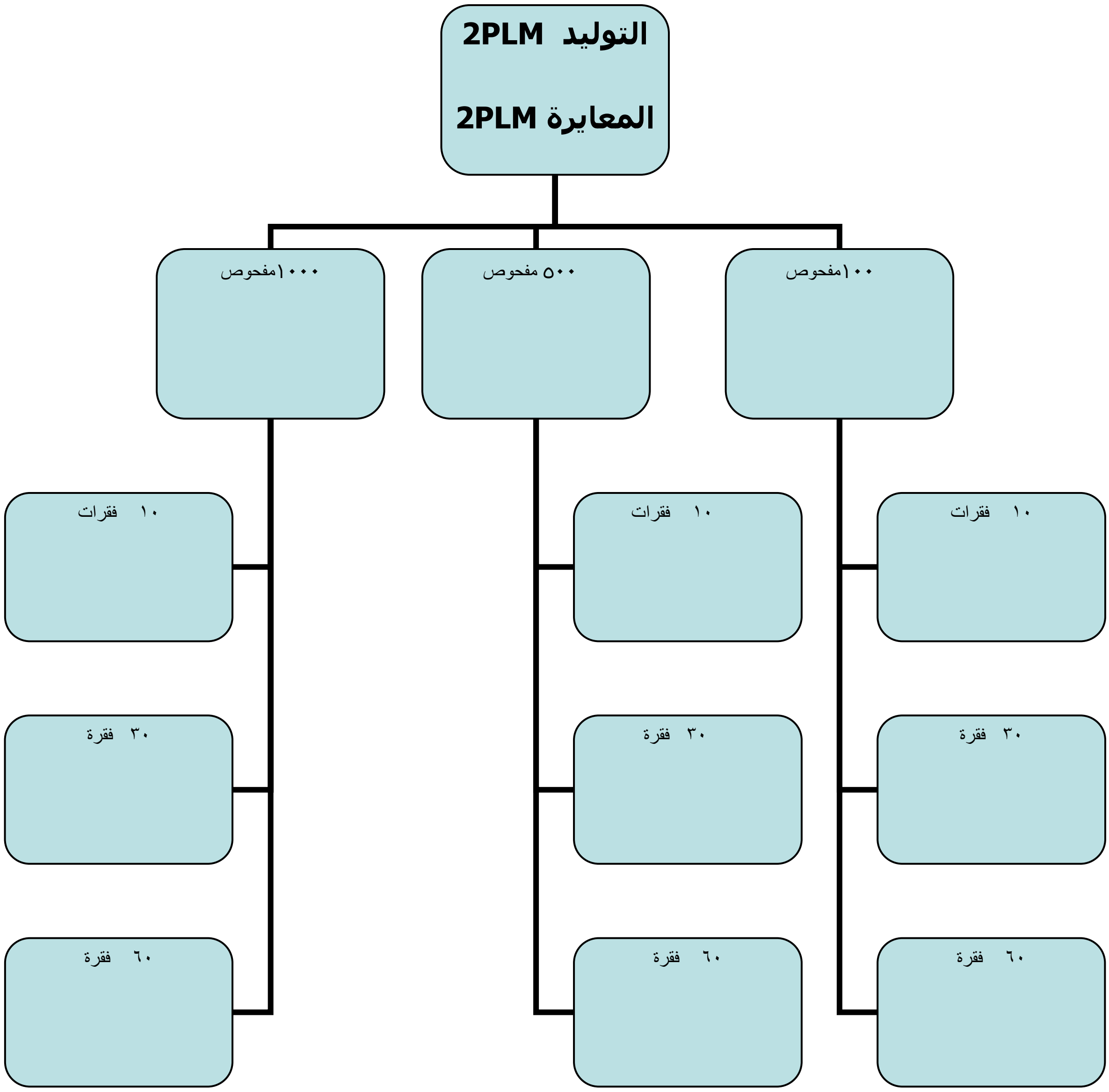
تم توليد استجابات ثنائية الإجابة تحت افتراضات النموذج ثنائي المعالم على ثلاثة مستويات من الطول (١٠، ٣٠، ٦٠) فقرة ،وثلاثة مستويات من حجم العينة (١٠٠، ٥٠٠، ١٠٠٠) مفحوص، ونتج تسعة خلايا مثلت كل الظروف المختلفة الناتجة عن تفاعلات مستويات طول الاختبار وحجم العينة تحت افتراضات النموذج ثنائي المعالم كما هو موضح بالشكل(١) ،حيث مثلت الخلية الأولى طول اختبار (١٠) فقرات و حجم عينة (١٠٠)مفحوص،و تم توليد بيانات الاستجابة على الفقرات لهذا الظرف بصعوبة تتوزع توزيعاً طبيعياً بمتوسط صفر وانحراف معياري واحد وبتوزيع منتظم لمعلم التمييز بمدى يتراوح من (١,٤ إلى ٢) ،و تتوزع قدرات عينة المفحوصين بتوزيع سوي بمتوسط صفر وانحراف معياري واحد، وتم إعادة توليد البيانات مئة مرة لينتج مئة مصفوفة من بيانات الاستجابة على الفقرات تحت هذا الظرف .وتم إعادة نفس الإجراء لجميع الخلايا الناتجة إذ نتج (٩٠٠) مصفوفة من بيانات الاستجابة على الفقرات ممثلة لكل الظروف .

ب- توليد البيانات المتعلقة بفحص الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة تحت افتراض النموذج اللوجستي ثلاثي المعالم.

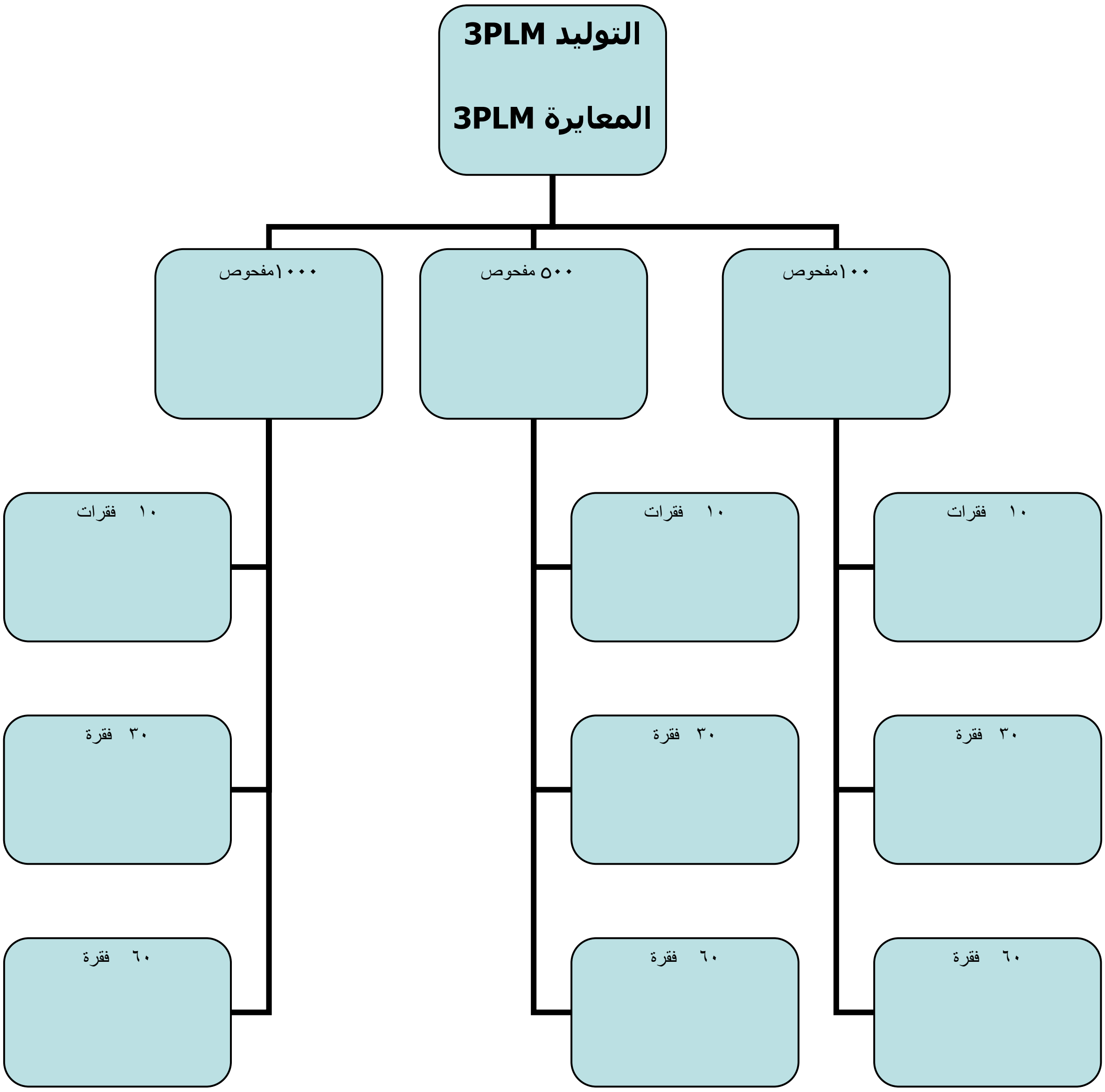
تم توليد استجابات ثنائية الإجابة تحت افتراضات النموذج ثلاثي المعالم على ثلاثة مستويات من الطول (١٠، ٣٠، ٦٠) فقرة ،وثلاثة مستويات من حجم العينة (١٠٠، ٥٠٠، ١٠٠٠) مفحوص، ونتج تسعة خلايا مثلت كل الظروف المختلفة الناتجة عن تفاعلات مستويات طول الاختبار وحجم العينة تحت افتراضات النموذج ثلاثي المعالم كما هو موضح بالشكل(٢) ،حيث مثلت الخلية الأولى طول اختبار (١٠) فقرات و حجم عينة (١٠٠)مفحوص،و تم توليد بيانات الاستجابة على الفقرات لهذا الظرف بصعوبة تتوزع توزيعاً طبيعياً بمتوسط صفر وانحراف معياري واحد ،وبتوزيع منتظم لمعلم التمييز بمدى يتراوح من (١,٤ إلى ٢)،وبتوزيع منتظم لمعلم التخمين بمدى يتراوح(٠,٢ إلى ٠,٣) ،و تتوزع قدرات عينة المفحوصين بتوزيع سوي بمتوسط صفر وانحراف معياري واحد،وتم إعادة توليد البيانات مئة مرة لينتج مئة مصفوفة من بيانات الاستجابة على الفقرات تحت هذا الظرف . وتم إعادة توليد البيانات مئة مرة لينتج مئة مصفوفة من بيانات الاستجابة على الفقرات تحت هذا الظرف .وتم إعادة نفس الإجراء لجميع الخلايا الناتجة إذ نتج (٩٠٠) مصفوفة من بيانات الاستجابة على الفقرات ممثلة لكل الظروف .

جـ - توليد البيانات المتعلقة بفحص قوة الاختبار لإحصائيات جودة مطابقة الفقـرة تحـت افتراض النموذج اللوجستي ثلاثي المعالم .

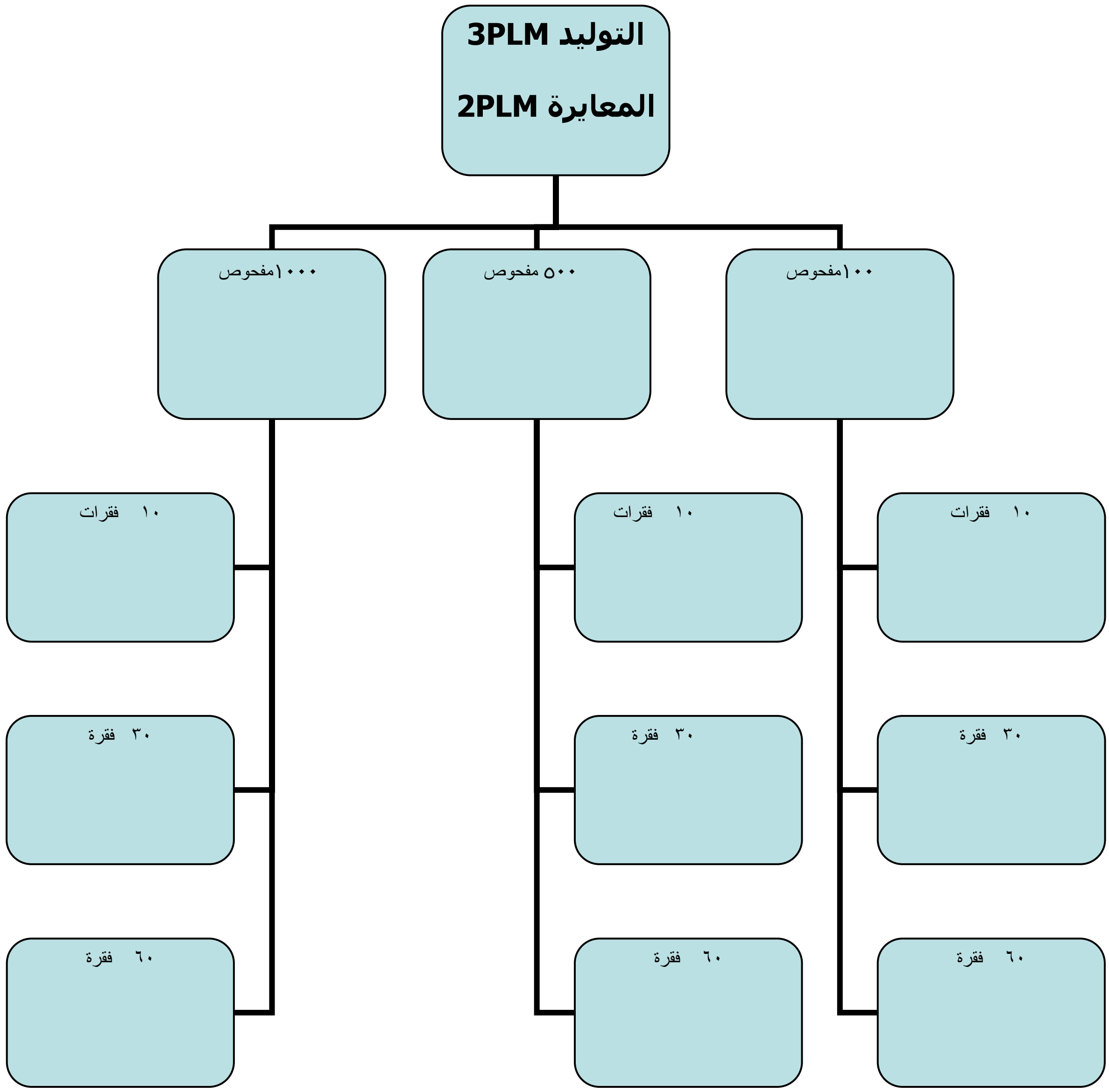
تم توليد استجابات ثنائية الإجابة تحت افتراضات النمـوذج ثلاثـي المعـالم بـنفس الظروف المذكورة سابقاً في الخطوة (ب)، ونتج ٩ خلايا ولكل ظرف في كل خلية تم إعـادة توليد البيانات مئة مرة لينتج مئة مصفوفة من بيانات الاستجابة على الفقرات. بحيـث ينـتج (٩٠٠) مصفوفة من بيانات الاستجابة على الفقرات ممثلة لكل الظروف والشكل (٣) يوضـح ذلك.



الشكل ١. مخطط بياني لفحص الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة تحت افتراض النموذج ثنائي المعالم



الشكل ٢. مخطط بياني لفحص الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة تحت افتراض النموذج ثلاثي المعالم



الشكل ٣. مخطط بياني لفحص قوة لاختبار لإحصائيات جودة مطابقة الفقرة

رابعا- طريقة فحص الخطأ من النوع الأول وقوة الاختبار لإحصائيات فحص جودة مطابقـة الفقرة:

أ. طريقة فحص الخطأ من النوع الأول :

الإستراتيجية التي اعتمدت للإجابة عن أسئلة الدراسة والمتعلقة بفحص الخطأ من النوع الأول، اســـتخدمت الطريقـــة التـــي اعتمـــدها أورلانـــدو وثيـــسن (Orlando and Thissen) والتي تقوم على أساس توليد البيانات بنموذج لوجستي معـين و تقدير المعالم و فحص إحصائيات جودة المطابقة باستخدام نفس النموذج ،والفرضية الصفرية تعتبر صحيحة لكل فقرات الاختبار ،حيث تم معايرة(Calibration) بيانات الاستجابات على الفقرات باستخدام نفس نموذج نظرية استجابة الفقرة الذي تم توليـدها تحـت افتراضـاته، و لفحص الخطأ من النوع الأول تحت افتراض النموذج اللوجستي ثنائي المعـالم لإحـصائيات جودة مطابقة الفقرة تم معايرة(Calibration) بيانات الاستجابات على الفقرات باستخدام نفس النموذج المستخدم في توليدها -هو النموذج اللوجستي ثنائي المعالم- و كـذلك تـم حـساب إحصائيات جودة مطابقة الفقرة من خلال البرامج الحاسوبية لكل إحصائي و التي تستخدم في هذا الظرف النموذج اللوجستي ثنائي المعالم في حساب قيم الإحصائيات ، و لفحص الخطـأ من النوع الأول تحت افتراض النموذج اللوجستي ثلاثي المعالم لإحصائيات جـودة مطابقـة الفقرة تم معايرة(Calibration) بيانات الاستجابات على الفقرات باسـتخدام نفـس النمـوذج المستخدم في توليدها -هو النموذج اللوجستي ثلاثي المعالم- و كذلك تم حساب إحـصائيات جودة مطابقة الفقرة من خلال البرامج الحاسوبية لكل إحصائي و التي تستخدم في هذا الظرف النموذج اللوجستي ثلاثي المعالم في حساب قيم الإحصائيات . ونظرياً فإن جميـع الفقـرات ستكون متطابقة لأنه تم توليد الاستجابات عليها ومعايرتها(Calibration) وفحـص جـودة المطابقة لها باستخدام نفس نموذج نظرية استجابة الفقرة ، وأي إحصائي من إحصائيات جودة مطابقة الفقرة يشير من خلال الدلالة الإحصائية المرتبطة بقيمته إلى أن هناك فروق يعتبـر هذا خطأ من النوع الأول ومعنى ذلك أن الإحصائي يكتشف فقرات غير مطابقة وهي في واقع الأمر مطابقة . وتم من خلال التالي:

بعد توليد بيانات الاستجابات باستخدام برنامج (Wingen2) تحت الظروف المختلفـة والتي تمثلت بتسعة خلايا وتتضمن كل خلية مئة مصفوفة من بيانات الاستجابات على فقرات الاختبار، تم تدريج الاستجابات المولدة باستخدام نفس النموذج الذي تم توليدها عليـه، سـواء

النموذج ثنائي المعالم أو ثلاثي المعالم. وتم ذلك باستخدام برامج حاسوبية هي Bilog، Bilog-Mg، لحساب ثلاثة إحصائيات لفحص جودة مطابقة الفقرة وهي:

١. إحصائي بك Bock (χ^2_B): استخدم برنامج Bilog لحساب هذا الإحصائي لكل فقرة في مصفوفات الاستجابات على الفقرات عند كل خلية و لجميع مرات الإعادة،إذ نتج (٩٠٠)مرة تحليل للإحصائي حيث تم حساب (χ^2_B) لكل فقرة في جميع مرات التحليل.وحددت مستويات الدلالة التالية لفحص إحصائيات جودة مطابقة الفقرة عندها وهي((٠,٠٥) و (٠,٠١) و (٠,١))،وبعد توليد استجابات البيانات على الفقرات وفحص جودة المطابقة للإحصائي بك في ظروف الدراسة المختلفة ،وعند كل ظرف حسبت عدد الحالات التي تساوت أو قلت في قيمتها عن مستوى الدلالة المحدد واعتبرت بأنها فقرات غير مطابقة للنموذج الرياضي و استخرجت تكرار و نسبة هذه الحالات من عدد الحالات الكلية وهي عدد الفقرات غير المطابقة في عدد مرات إعادة توليد البيانات ،واعتبرت هذه النسبة الخطأ من النوع الأول تحت توزيع المعاينة المحتمل لتوزيع الإحصائي.

٢. إحصائي الأرجحية العظمى (G^2): استخدم برنامج (Bilog-Mg) لحساب هذا الإحصائي لكل فقرة في مصفوفات الاستجابات على الفقرات عند كل خلية و لجميع مرات الإعادة،إذ نتج (٩٠٠)مرة تحليل للإحصائي حيث تم حساب (G^2) لكل فقرة في جميع مرات التحليل.وحددت مستويات الدلالة التالية لفحص إحصائيات جودة مطابقة الفقرة عندها وهي((٠,٠٥) و (٠,٠١) و (٠,١))،وبعد توليد استجابات البيانات على الفقرات وفحص جودة المطابقة للإحصائي الأرجحية العظمى في ظروف الدراسة المختلفة ،وعند كل ظرف حسبت عدد الحالات التي تساوت أو قلت في قيمتها عن مستوى الدلالة المحدد واعتبرت بأنها فقرات غير مطابقة للنموذج الرياضي و استخرجت تكرار و نسبة هذه الحالات من عدد الحالات الكلية وهي عدد الفقرات غير المطابقة في عدد مرات إعادة توليد البيانات ،واعتبرت هذه النسبة الخطأ من النوع الأول تحت توزيع المعاينة المحتمل لتوزيع الإحصائي.

٣. إحصائي البواقي المعيارية (RMS): استخدم برنامج Bilog لحساب هذا الإحصائي لكل فقرة في مصفوفات الاستجابات على الفقرات عند كل خلية و لجميع مرات الإعادة،إذ نتج (٩٠٠)مرة تحليل للإحصائي حيث تم حساب (RMS) لكل فقرة في جميع مرات التحليل.وحددت مستويات الدلالة التالية لفحص إحصائيات جودة مطابقة الفقرة عند قيم Z المطلقة وهي[Z=١ ، Z=٢ ، Z=٣]،وبعد توليد استجابات البيانات على الفقرات

،وفحص جودة المطابقة للإحصائي (RMS) في ظروف الدراسة المختلفة ،وعند كل ظرف حسبت عدد الحالات التي تساوت أو قلت في قيمتها عن مستوى الدلالة المحدد واعتبرت بأنها فقرات غير مطابقة للنموذج الرياضي و استخرجت تكرار و نسبة هذه الحالات من عدد الحالات الكلية وهي عدد الفقرات غير المطابقة في عدد مرات إعادة توليد البيانات ،واعتبرت هذه النسبة الخطأ من النوع الأول تحت توزيع المعاينة المحتمل لتوزيع الإحصائي.

ب. طريقة فحص قوة الاختبار لإحصائيات جودة مطابقة الفقرة:

الإستراتيجية التي أُعتمدت للإجابة عن أسئلة الدراسة والمتعلقة بفحص قوة الاختبار لإحصائيات جودة مطابقة الفقرة،استخدمت الطريقة التي اعتمدها أورلاندو وثيسن (Orlando and Thissen) والتي تقوم على أساس توليد البيانات بنموذج لوجستي معين و تقدير المعالم و فحص إحصائيات جودة المطابقة باستخدام نموذج مختلف والفرضية الصفرية تعتبر خاطئة لكل فقرات الاختبار،وفي الدراسة الحالية تم توليد استجابات على الفقرات تحت افتراضات النموذج ثلاثي المعالم اعتماداً على الظروف المذكورة سابقا في الخطوة (ج) في مرحلة توليد البيانات، ومعايرتها(Calibration) باستخدام نموذج رياضي مختلف عن النموذج المستخدم بتوليد البيانات وهو النموذج ثنائي المعالم .وتم حساب إحصائيات جودة مطابقة الفقرة تحت افتراض النموذج ثنائي المعالم، ونظرياً ستكون جميع الفقرات غير مطابقة لأنها تمت معايرتها على نموذج مختلف عن النموذج الذي استخدم في توليد البيانات ، وقوة الاختبار تتمثل في قدرة الإحصائي على كشف الفقرات غير المطابقة. وأي إحصائي من إحصائيات جودة مطابقة الفقرة يشير من خلال الدلالة الإحصائية المرتبطة بقيمته إلى أن هناك فروق، يعتبر هذا مؤشر على قوة الاختبار في كشف الفقرات غير المطابقة. ومعنى ذلك أن الإحصائي يكتشف فقرات غير مطابقة وهي فعلا غير مطابقة.و بعد توليد بيانات الاستجابات باستخدام برنامج (Wingen2) تحت الظروف المختلفة والتي تمثلت بتسعة خلايا إذ تتضمن كل خلية مئة مصفوفة من بيانات الاستجابات على فقرات الاختبار، تم معايرة الاستجابات المولدة باستخدام نموذج يختلف عن النموذج الذي تم توليدها عليه. وتم ذلك باستخدام برامج حاسوبية هي Bilog، Bilog-Mg، لحساب ثلاثة إحصائيات لفحص جودة مطابقة الفقرة وهي:

١. إحصائي بك Bock (χ^2_B): استخدم برنامج Bilog لحساب هذا الإحصائي لكل فقرة في مصفوفات الاستجابات على الفقرات عند كل خلية و لجميع مرات الإعادة،إذ نتج (٩٠٠)مرة تحليل للإحصائي حيث تم حساب (χ^2_B) لكل

فقرة في جميع مرات التحليل.وحددت مـستويات الدلالـة التاليـة لفحـص إحصائيات جودة مطابقة الفقرة عنـدها وهـي((٠,٠٥) و (٠,٠١) و (٠,١))،وبعد توليد استجابات البيانات على الفقـرات وفحـص جـودة المطابقـة للإحصائي بك في ظروف الدراسة المختلفة ،وعند كل ظرف حـسبت عـدد الحالات التي قلت في قيمتها عن مستوى الدلالة المحدد واعتبرت بأنها فقرات غير مطابقة للنموذج الرياضي و استخرجت تكرار و نسبة هذه الحالات مـن عدد الحالات الكلية وهي عدد الفقرات غير المطابقة في عدد مـرات إعـادة توليد البيانات ،واعتبرت هذه النسبة قوة الاختبـار تحـت توزيـع المعاينـة المحتمل لتوزيع الإحصائي.

٢. إحصائي الأرجحية العظمى (G^2): استخدم برنامج (Bilog-Mg) لحـساب الإحصائي هذا الإحصائي لكل فقرة في مصفوفات الاستجابات على الفقرات عند كل خلية و لجميع مرات الإعادة،إذ نتج (٩٠٠)مرة تحليـل للإحـصائي حيث تم حساب (G^2) لكل فقرة في جميع مرات التحليل.وحـددت مـستويات الدلالة التالية لفحص إحصائيات جودة مطابقة الفقرة عندها وهـي((٠,٠٥) و (٠,٠١) و (٠,١))،وبعد توليد استجابات البيانات على الفقرات وفحص جودة المطابقة للإحصائي الأرجحية العظمى في ظروف الدراسة المختلفة ،وعنـد كل ظرف حسبت عدد الحالات التي قلت في قيمتها عن مستوى الدلالة المحدد واعتبرت بأنها فقرات غير مطابقة للنموذج الرياضي و استخرجت تكـرار و نسبة هذه الحالات من عدد الحالات الكلية وهي عدد الفقرات غير المطابقـة من عدد مرات إعادة توليد البيانات ،واعتبرت هذه النسبة قوة الاختبار تحـت توزيع المعاينة المحتمل لتوزيع الإحصائي.

٣. إحصائي البواقي المعيارية (RMS): استخدم برنامج Bilog لحـساب هـذا الإحصائي لكل فقرة في مصفوفات الاستجابات على الفقرات عند كل خلية و لجميع مرات الإعادة،إذ نتج (٩٠٠)مرة تحليل للإحصائي حيث تـم حـساب (RMS) لكل فقرة في جميع مرات التحليل.وحددت مستويات الدلالة التاليـة لفحص إحصائيات جودة مطابقة الفقرة عند قيم Z المطلقة وهـي [Z=١ ، Z=٢ ، Z=٣]،وبعد توليد استجابات البيانات على الفقرات ،وفحص جـودة المطابقة للإحصائي (RMS) في ظروف الدراسة المختلفة ،وعند كل ظرف

حسبت عدد الحالات التي قلت في قيمتها عن مستوى الدلالة المحدد واعتبرت بأنها فقرات غير مطابقة للنموذج الرياضي و استخرجت تكرار و نسبة هـذه الحالات من عدد الحالات الكلية وهي عدد الفقرات غير المطابقة فـي عـدد مرات إعادة توليد البيانات ،واعتبرت هذه النسبة قوة الاختبار تحـت توزيـع المعاينة المحتمل لتوزيع الإحصائي.

المعالجات الإحصائية:

تم تحليل الفقرات بهدف الحصول على محكات إحصائية للدلالة على مطابقة الفقرات لتوقعات النموذج في نظرية استجابة الفقرة حيث استخرج:

1. تم استخراج قيم إحصائيات فحص جودة مطابقة الفقرة الـثلاث واحتماليتهـا بالنـسبة لدرجات الحرية لكل فقرة.
2. حساب الخطأ من النوع الأول، إذ أن الوضع الأمثل حتى تكون الفقرة مطابقة لتوقعـات النموذج، أن تكون قيمة إحصائي مطابقة الفقرة بالنسبة لدرجات الحريـة ليـست ذات الدلالة. مما يعني أن الفروق بين البيانات المشاهدة والنموذج ليس دالة.
3. حساب قوة الاختبار لإحصائيات جودة مطابقة الفقرة إذ أن الوضع الأمثل حتى تكـون الفقرة غير مطابقة لتوقعات النموذج أن تكون قيمة إحصائي مطابقـة الفقـرة بالنـسبة لدرجات الحرية ذات الدلالة، ما يعني أن الفروق بين البيانـات المـشاهدة والمتوقعـة للنموذج دالة.
4. استخدام الجداول والرسومات البيانية التي توضح تأثير متغيرات الدراسة على الخطـأ من النوع الأول وقوة الاختبار لإحصائيات جودة مطابقة الفقرة.

الفصل الرابع

النتائج

الفصل الرابع

النتائج

هدفت الدراسة إلى فحص الخطأ من النوع الأول، وقوة الاختبار لإحصائيات جـودة مطابقة الفقرة تحت ظروف نوع النموذج الرياضي وطول الاختبار وحجم العينة وتفاعلاتهـا وبالتحديد حاولت الدراسة الإجابة عن الأسئلة التالية:

١. هل يختلف الخطأ من النوع الأول وقوة الاختبار لإحصائيات فحص جـودة مطابقـة الفقرة (إحصائي بك (χ^2_B)، وإحصائي (G^2) للأرجحية العظمى وإحصائي البـواقي المعيارية (RMS))، باختلاف طول الاختبار ؟

٢. هل يختلف الخطأ من النوع الأول و قوة الاختبار لإحصائيات فحص جودة مطابقـة الفقرة (إحصائي بك (χ^2_B)، وإحصائي (G^2) للأرجحية العظمى وإحصائي البـواقي المعيارية (RMS))، باختلاف حجم العينة ؟

٣. هل يختلف الخطأ من النوع الأول لإحصائيات فحص جودة مطابقة الفقرة (إحـصائي بك (χ^2_B)، وإحصائي (G^2) للأرجحيـة العظمـى وإحـصائي البـواقي المعياريـة (RMS))، باختلاف النموذج الرياضي المستخدم ؟

٤. هل يختلف الخطأ من النوع الأول لإحصائيات فحص جودة مطابقة الفقرة (إحـصائي بك (χ^2_B)، وإحصائي (G^2) للأرجحيـة العظمـى وإحـصائي البـواقي المعياريـة (RMS))، باختلاف التفاعل بين متغيرات: طول الاختبار وحجم العينة؟

٥. هل تختلف قوة الاختبار الإحصائي لإحصائيات فحص جودة مطابقة الفقرة (إحصائي بك (χ^2_B)، وإحصائي (G^2) للأرجحيـة العظمـى وإحـصائي البـواقي المعياريـة (RMS))، باختلاف التفاعل بين متغيرات: طول الاختبار وحجم العينة ؟

وفيما يلي عرضا للنتائج:

- النتائج المتعلقة بالسؤال الأول " هل تختلف الخطأ مـن النـوع الأول و قـوة الاختبـار الإحصائي لإحصائيات فحص جودة مطابقة الفقرة باختلاف طول الاختبار ؟"

للإجابة على الجزء الذي يتعلق بمعرفة أثر طول الاختبار على الخطأ مـن النـوع الأول لإحصائيات جودة مطابقة الفقرة عند ثبات حجم العينة، ولدت الفقرات بالآلية المـذكورة سابقا ،و استخراج توزيع المعاينة لإحصائيات جودة المطابقة ومن ثم حسب الخطأ من النوع

الأول عند كل ظرف من ظروف الاختبار الثلاثة بمستويات الدلالـة المحـددة وهـي(٠,١ ، ٠,٠٥ ، ٠,٠١) ،و يمثل الخطأ من النوع الأول معدل عدد الحالات غير المطابقـة ،وذلـك بحساب نسبة هذه الحالات غير المطابقة من عدد الحالات المولدة ،من خـلال قـسمة عـدد الفقرات غير المطابقة على عدد فقرات الاختبار مضروبا في عدد مرات الإعـادة ، فعلـى سبيل المثال: الخلية الأولى في الجدول (٣) عند حجم عينة ١٠٠ مفحوص وطول اختبار ١٠ فقرات كان الخطأ من النوع الأول للإحصائي (G^2) عند مستوى دلالة(٠,١) هـو (٠,١٣٨) .ويمكن تلخيص العملية الحسابية بالخطوات التالية:

١-توليد بيانات لاختبار بطول ١٠ فقرات وتم إعادة التوليد مائة مرة.

٢-حساب قيمة الإحصائي(G^2) ودلالاتها الإحصائية لكل فقرة من الفقرات الناتجة في عـدد المرات المولدة ،حيث بلغ عددها في جميع مرات التوليد (١٠٠٠)فقرة .

٣- مثل توزيع المعاينة لقيم الإحصائي(G^2) وحسبت عدد الحالات التي قلت أو تساوي فـي قيمتها عن مستوى الدلالة(٠,١) واعتبرت بأنها فقرات غير مطابقة للنموذج ،حيث كـان عددها (١٣٨)فقرة من أصل (١٠٠٠)فقرة .

٤-حسبت النسبة تحت منحى توزيع المعاينة و التي تمثل الخطأ من النوع الأول بإيجاد نسبة عدد الفقرات غير المطابقة عبر مائة مرة إعادة مقسوما على مضروب طول الاختبار في عدد مرات الإعادة.

الخطأ من النوع الأول=١٣٨/(١٠*١٠٠)=٠,١٣٨

ولمزيد من الإيضاح يمكن تفصيل القيمة الواردة في الخلية الثانية في الجدول (٣) عند حجم عينة ١٠٠ ومفحوص وطول اختبار ٣٠ فقرة ،كان الخطأ من النوع الأول للإحـصائي (G^2) عند مستوى دلالة(٠,١) هو (٠,٠٧٩) . ويمكن تلخيص العملية الحسابية بـالخطوات التالية:

١-توليد بيانات لاختبار بطول ٣٠ فقرة وتم إعادة التوليد مائة مرة.

٢-حساب قيمة الإحصائي(G^2) ودلالاتها الإحصائية لكل فقرة من الفقرات الناتجة في عـدد المرات المولدة ،حيث بلغ عددها في جميع مرات التوليد (٣٠٠٠)فقرة .

٣- مثل توزيع المعاينة لقيم الإحصائي(G^2) وحسبت عدد الحالات التي قلت أو تساوي فـي قيمتها عن مستوى الدلالة(٠,١) واعتبرت بأنها فقرات غير مطابقة للنموذج،حيـث كـان عددها (٢٣٩)فقرة من أصل (٣٠٠٠)فقرة .

٤- حسبت النسبة تحت منحى توزيع المعاينة و التي تمثل الخطأ من النـوع الأول بإيجـاد نسبة عدد الفقرات غير المطابقة عبر مائة مرة إعادة مقـسوما علـى مـضروب طـول الاختبار في عدد مرات الإعادة.و الخطأ من النوع الأول=٢٣٩/(٣٠*١٠٠)=٠,٠٧٩

وكذلك الحال للخلية الثالثة في الجدول (٣) عند حجم عينة ١٠٠ ومفحوص وطـول اختبـار ٦٠ فقرات كان الخطأ من النوع الأول للإحصائي (G^2)عنـد مـستوى دلالـة(٠,١) هـو (٠,٠١٦) .وهذه القيمة كانت ناتج عن الخطوات التالية:

١-توليد بيانات لاختبار بطول ٦٠ فقرات وتم إعادة التوليد مائة مرة.

٢-حساب قيمة الإحصائي(G^2) ودلالاتها الإحصائية لكل فقرة من الفقرات الناتجة في عـدد المرات المولدة ،حيث بلغ عددها في جميع مرات التوليد (٦٠٠٠)فقرة .

٣- مثل توزيع المعاينة لقيم الإحصائي(G^2) وحسبت عدد الحالات التي قلت أو تساوي فـي قيمتها عن مستوى الدلالة(٠,١) واعتبرت بأنها فقرات غير مطابقة للنموذج ،حيث كان عددها (٩٨)فقرة من أصل (٦٠٠٠)فقرة .

٤- حسبت النسبة تحت منحى توزيع المعاينة و التي تمثل الخطأ من النـوع الأول بإيجـاد نسبة عدد الفقرات غير المطابقة عبر مائة مرة إعادة مقسوما علـى مـضروب طـول الاختبار في عدد مرات الإعادة.

الخطأ من النوع الأول=٩٨/(٦٠*١٠٠)=٠,٠١٦

الجدول ٣. الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة تحت افتراض النموذج اللوجستي ثنائي المعالم تبعاً لاختلاف طول الاختبار

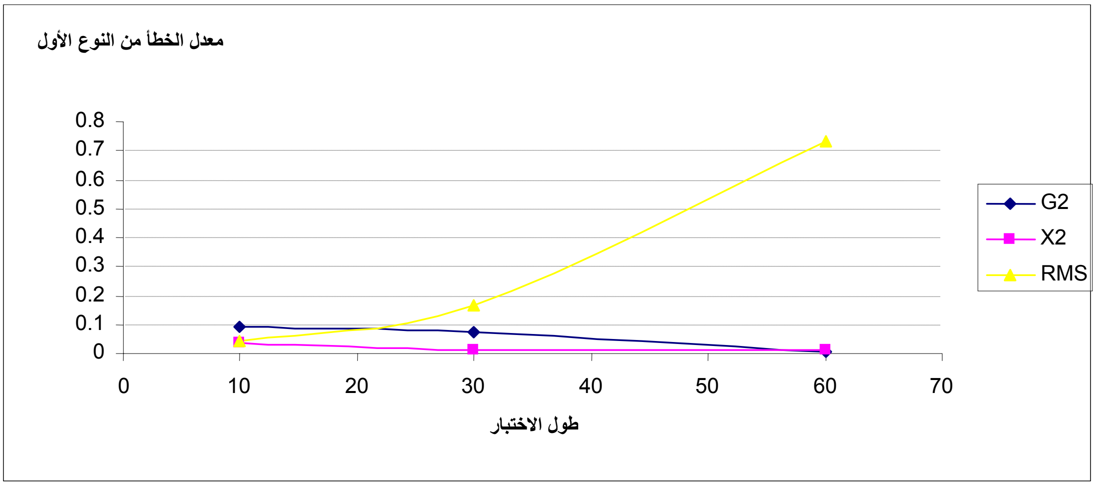
حجم العينة	طول الاختبار (عدد الفقرات)	إحصائي جودة مطابقة الفقرة								
		إحصائي (G^2) الأرجحية العظمى.			إحصائي Bocks (χ^2_B)			إحصائي البواقي المعيارية (RMS)		
		مستوى الدلالة(α)			مستوى الدلالة(α)			مستوى الدلالة Z		
		٠,١	٠,٠٥	٠,٠١	٠,١	٠,٠٥	٠,٠١	Z=١	Z=٢	Z=٣
١٠٠	١٠	١٣٨ (٠,١٣٨)	٩٠ (٠,٠٩)	٥٨ (٠,٠٥٨)	٧٥ (٠,٠٧٥)	٣٨ (٠,٠٣٨)	١٣ (٠,٠١٣)	٥٠٣ (٠,٥٠٣)	٤٦ (٠,٠٤٦)	١١ (٠,٠١١)
	٣٠	٢٣٩ (٠,٠٧٩)	٢١٦ (٠,٠٧٢)	١٧٨ (٠,٠٥٩)	٩٠ (٠,٠٣)	٤١ (٠,٠٤١)	٥ (٠,٠٠٢)	١٦١٢ (٠,٥٣٧)	٤٩٥ (٠,١٦٥)	٢٠٨ (٠,٠٦٩)
	٦٠	٩٨ (٠,٠١٦)	٣٥ (٠,٠٠٦)	٥ (٠,٠٠٠٨)	١٥٤ (٠,٠٢٦)	٧٠ (٠,٠١٢)	١٢ (٠,٠٠٢)	٥٨٩١ (٠,٩٨٢)	٤٣٨٩ (٠,٧٣٢)	٢٠٨٧ (٠,٣٤٨)
٥٠٠	١٠	٢١٩ (٠,٢١٩)	١٥١ (٠,١٥١)	٦٦ (٠,٠٦٦)	٢٨١ (٠,٢٨١)	١٨٢ (٠,١٨٢)	٦٧ (٠,٠٦٧)	٢٢٥ (٠,٢٢٥)	٧ (٠,٠٠٧)	صفر (صفر)
	٣٠	١١٤ (٠,٠٣٨)	٥٥ (٠,٠١٨)	١٤ (٠,٠٠٥)	٤٠٣ (٠,١٣٤)	٢١٦ (٠,٠٧٢)	٥٣ (٠,٠١٨)	٢٤٥٥ (٠,٨١٨)	٨١٣ (٠,٢٧)	١٧٤ (٠,٠٥٨)
	٦٠	١٧٢ (٠,٠٢٩)	٧٣ (٠,٠١٢)	١١ (٠,٠٠٢)	٣٣٩ (٠,٠٥٧)	١٦٣ (٠,٠٢٧)	٢٧ (٠,٠٠٥)	٥٩٣٠ (٠,٩٩)	٣٧٩٠ (٠,٦٣)	١٢٦٤ (٠,٢١)
١٠٠٠	١٠	٧٧٤ (٠,٧٧٤)	٦١٣ (٠,٦١٣)	٣٢٣ (٠,٣٢٣)	٨٠٧ (٠,٨٠٧)	٦٨٥ (٠,٦٨٥)	٤٤٣ (٠,٤٤٣)	٩٦ (٠,٠٩٦)	٤ (٠,٠٠٤)	صفر (صفر)
	٣٠	١٨٣ (٠,٦١)	٨٩ (٠,٢٩)	٩ (٠,٠٠٣)	٦٣٨ (٠,٢١٣)	٣٧٣ (٠,١٢٤)	٩٥ (٠,٠٣٢)	٢٣٧٤ (٠,٧٩)	٦٢٥ (٠,٢١)	٧٥ (٠,٠٣)
	٦٠	٢٣٨ (٠,٠٣٩)	١٠٨ (٠,٠١٨)	١٧ (٠,٠٠٣)	٣٣٨ (٠,٠٥٦)	١٥٦ (٠,٠٢٦)	٢٦ (٠,٠٠٤)	٥٩٢٩ (٠,٠٠٤)	٣٢٦٠ (٠,٥٤)	٦٢٣ (٠,١٠٤)

ويبين الجدول (٣) الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة وهي إحصائي الأرجحية العظمى (G^2-index) وإحصائي بك (Bock-index) (χ^2_B) وإحصائي البواقي المعيارية (RMS) ،وتمثل القيم في الجدول المساحة تحت منحى توزيع المعاينة و التي تعكس التكرارات والنسبة المئوية لعدد الفقرات غير المطابقة عبر مائة مرة إعادة، حيث تم توليد الاستجابات على الفقرات ومعايرتها(Calibration) تحت افتراض النموذج ثنائي المعالم.

ويتضح من الجدول(٣):

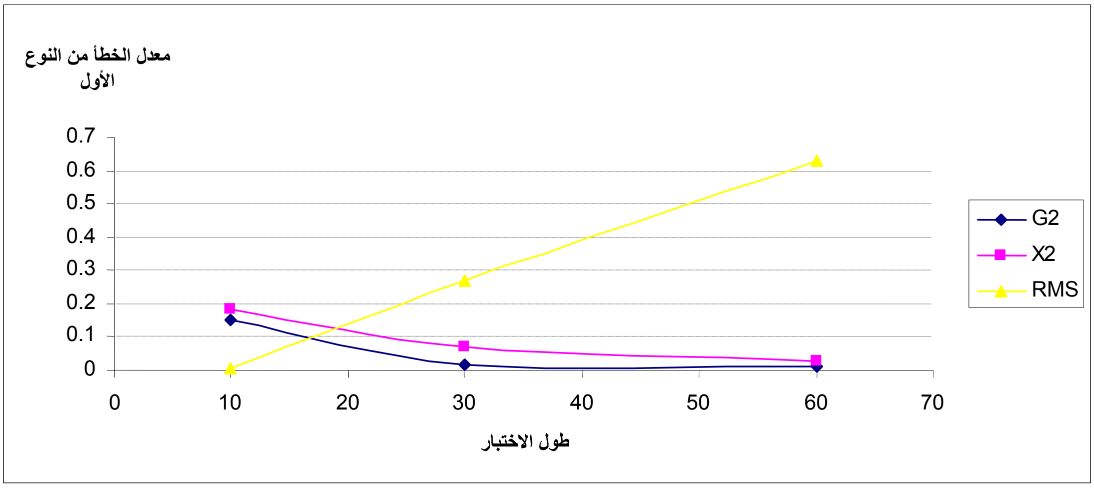
* أنه بزيادة طول الاختبار ونقصان قيمة مستوى الدلالة الإحصائية يقل الخطأ من النوع الأول، لإحصائيات جودة مطابقة الفقرة (G^2-index) و (χ^2_B) عند حجم عينة (١٠٠) مفحوص، في حين أن الخطأ من النوع الأول يزيد مع زيادة طول الاختبار لإحصائي جودة مطابقة الفقرة "البواقي المعيارية (RMS)" عند نفس حجم العينة وعند جميع مستويات الدلالة الإحصائية.
* ونلاحظ عند مقارنة إحصائيات جودة مطابقة الفقرة عند حجم عينة (١٠٠) مفحوص وطول اختبار بعشرة فقرات، تحت افتراض النموذج ثنائي المعالم الذي استخدم بتوليد والمعايرة، نلاحظ بأن الإحصائي بك (χ^2_B) يتمتع بخطأ من النوع الأول أقل مقارنة بالإحصائي (G^2)، في حين أبدى الإحصائي (البواقي المعيارية) (RMS) تحسناً كبيراً في الأداء وبخطأ أقل من الإحصائيين الآخرين عند مستوى دلالة Z=٣ عند طول اختبار (١٠) فقرات، وكان الأسوأ في الأداء مقارنة بالإحصائيين الآخرين عند طول ٣٠ فقرة و٦٠ فقرة.
* وكذلك نلاحظ عند مقارنة إحصائيات جودة مطابقة الفقرة عند حجم عينة (١٠٠) مفحوص، وطول اختبار (٣٠) فقرة، تحت نفس الظروف السابقة، نلاحظ بأن إحصائي جودة مطابقة الفقرة (χ^2_B) يأتي بالمرتبة الأولى لتمتعه بخطأ من النوع الأول أقل عند جميع مستويات الدلالة، مقارنة بالإحصائيين الآخرين، ويأتي بالمرتبة الثانية الإحصائي (G^2) وكان الإحصائي (RMS) الأسوأ في الأداء مقارنة بالإحصائيين الآخرين.
* وعند المقارنة عند طول اختبار (٦٠) فقرة وتحت نفس الظروف نلاحظ بأن إحصائي جودة المطابقة (G^2) أبدى معدلات خطأ من النوع الأول أقل عند جميع مستويات الدلالة مقارنة بالإحصائيين الآخرين وجاء الإحصائي (χ^2_B) بالمرتبة الثانية في حين أن

الإحصائي (RMS) جاء بخطأ من النوع الأول مرتفعة مقارنة بالإحـصائيين الآخـرين والشكل (٤) يوضح ذلك.

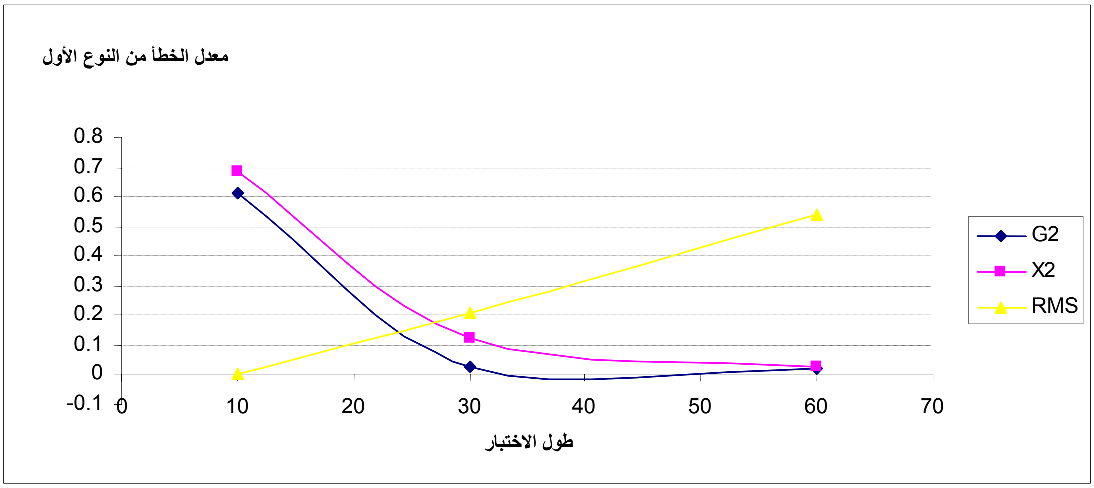


الشكل ٤. الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة (عند حجم عينة ١٠٠ فرد) ومستوى دلالة (٠,٠٥) تحت افتراض النموذج اللوجستي ثنائي المعالم

- وعند حجم عينة (٥٠٠) و(١٠٠٠) مفحوص كما يبين الجـدول (٣)، كانـت نتـائج أداء الإحصائيات(χ^2_B) و(G^2) مماثلة للظرف السابق، ونلاحظ أن الخطأ مـن النـوع الأول لإحصائيات جودة مطابقة الفقرة (χ^2_B) و(G^2) تقل مع زيادة طول الاختبـار ونقـصان مستوى الدلالة الإحصائية عند جميع مستويات طول الاختبار ، وكما توضحها الاشـكال (٥) و(٦) في حين يبدي الإحصائي (RMS) سلوكا معاكسا، وذلك بزيـادة الخطـأ مـن النوع الأول بزيادة طول الاختبار، ونجد أن الإحصائي (RMS) يتمتع بخطأ من النـوع الأول منخفضة تصل لصفر عند الاختبارات القصيرة بطول (١٠) فقرات. وكان ذو الأداء الأفضل من الإحصائيين الآخرين عند هذه الظروف ولكن بزيادة طول الاختبار فـإن أداء هذا الإحصائي ينخفض بشكل حاد وتمييز بزيادة الخطأ من النوع الأول عند زيادة طـول الاختبار إلى (٣٠) فقرة و(٦٠) فقرة، في حين يتحسن أداء كل من الإحصائيين الآخـرين مع زيادة طول الاختبار ويتبين أن الإحصائي(G) يتمتع بخطأ من النوع الأول أقل عنـد جميع مستويات الدلالة مقارنة بالآخرين تحت طول اختبـار (٣٠) و(٦٠) فقـرة ويـأتي بالمرتبة الثانية إحصائي (χ^2_B).



الشكل ٥. الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة (عند حجم عينة ٥٠٠فرد) ومستوى دلالة (٠,٠٥) تحت افتراض النموذج اللوجستي ثنائي المعالم



الشكل ٦. الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة (عند حجم عينة ١٠٠٠فرد) ومستوى دلالة (٠,٠٥) تحت افتراض النموذج اللوجستي ثنائي المعالم

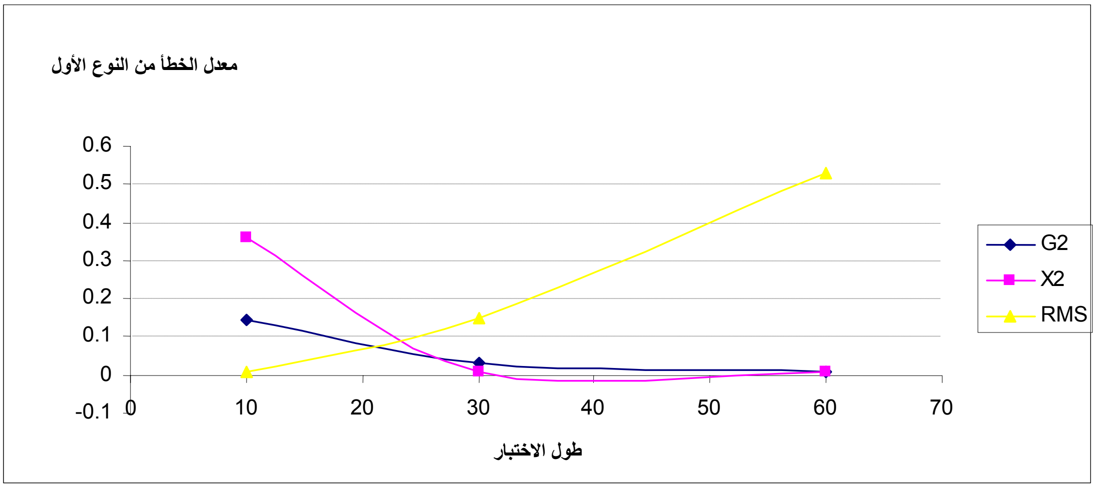
الجدول ٤. الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة تحت افتراض النموذج اللوجستي ثلاثي المعالم تبعاً لاختلاف طول الاختبار

حجم العينة	طول الاختبار (عدد الفقرات)	إحصائي جودة مطابقة الفقرة								
		إحصائي (G^2) الأرجحية العظمى.			إحصائي Bocks (χ^2_B)			إحصائي البواقي المعيارية (RMS)		
		مستوى الدلالة(α)			مستوى الدلالة(α)			مستوى الدلالة Z		
		٠,١	٠,٠٥	٠,٠١	٠,١	٠,٠٥	٠,٠١	Z=١	Z=٢	Z=٣
١٠٠	١٠	١٨٢ (٠,١٨٣)	١٤٣ (٠,١٤٣)	٩٦ (٠,٠٩٦)	٤٩٤ (٠,٤٩٤)	٣٦٤ (٠,٣٦)	١٥٧ (٠,١٥٧)	١٨٠ (٠,١٨)	٧ (٠,٠٠٧)	صفر (صفر)
	٣٠	١٩٧ (٠,٠٦٦)	٨٥ (٠,٠٣)	٤ (٠,٠٠١)	١٠٢ (٠,٠٣)	٣٩ (٠,٠١)	١ (٠,٠٠٠٣)	٢٠٦٤ (٠,٦٩)	٤٥٨ (٠,١٥)	١١٤ (٠,٠٣٨)
	٦٠	١٤٧ (٠,٠٣)	٧٩ (٠,٠١)	٤٥ (٠,٠٠٨)	١٩٨ (٠,٠٣)	٨٥ (٠,٠١)	٥ (٠,٠٠٠٨)	٥٨٥٦ (٠,٩٨)	٣١٩٣ (٠,٥٣)	٩٩٨ (٠,١٦)
٥٠٠	١٠	٥٥٥ (٠,٥٦)	٤٥١ (٠,٤٥)	٣٠٥ (٠,٣)	٩٩٠ (٠,٩٩)	٩٨٨ (٠,٩٩)	٩٨٦ (٠,٩٩)	٢٣٣ (٠,٢٣٣)	١٢ (٠,٠١٢)	٢ (٠,٠٠٢)
	٣٠	١٤٤ (٠,٠٥)	٦٣ (٠,٠٢)	١٠ (٠,٠٠٣)	٢٦٠ (٠,٠٩)	١٢٧ (٠,٠٤)	١٩ (٠,٠٠٦)	١٤٤٦ (٠,٤٨)	١٢٣ (٠,٠٤١)	٢٣ (٠,٠٠٨)
	٦٠	١٩٠ (٠,٠٣)	٧٨ (٠,٠١)	١٠ (٠,٠٠٢)	٢٩٣ (٠,٠٥)	١٢٩ (٠,٠٢)	١٩ (٠,٠٠٣)	٥٤٣٣ (٠,٩١)	١٢٥٠ (٠,٢١)	١١٧ (٠,٠٢)
١٠٠٠	١٠	٨٥٩ (٠,٨٦)	٧٦٨ (٠,٧٧)	٥٦٩ (٠,٥٧)	٩٦٣ (٠,٩٦)	٩٦٣ (٠,٩٦)	٩٦٢ (٠,٩٦)	٥٣ (٠,٠٥٣)	صفر (صفر)	صفر (صفر)
	٣٠	٣٨١ (٠,١٣)	٢٣٣ (٠,٠٧٨)	٨١ (٠,٠٣)	٥٥٠ (٠,١٨)	٣٤٩ (٠,١٢)	١٠٣ (٠,٠٣)	٢٢٥٣ (٠,٧٥)	٧٢ (٠,٠٢)	٧ (٠,٠٠٢)
	٦٠	٢٤٦ (٠,٠٤)	٩٣ (٠,٠٢)	١٤ (٠,٠٠٢)	٢٩٧ ٠,٠٥	١٣٧ (٠,٠٢)	٢٦ (٠,٠٠٤)	٥٢٤٦ (٠,٨٧)	٨٨٣ (٠,١٥)	٤٤ (٠,٠٠٧)

ويبين الجدول (٤): الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة وهـي إحصائي الأرجحية العظمى (G^2-index) وإحصائي بك (Bock-index) (χ^2_B) وإحـصائي البواقي المعيارية (RMS) ،وتمثل القيم في الجدول المساحة تحت منحى توزيـع المعاينـة و التي تعكس التكرارات والنسبة المئوية لعدد الفقرات غير المطابقة عبر مائة مرة إعادة، حيث تم توليد الاستجابات على الفقرات ومعايرتها(Calibration) تحت افتراض النمـوذج ثلاثـي المعالم.

ويتبين من الجدول (٤) :

* أن الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقـرة (G^2) و (χ^2_B) عنـد حجم عينة (١٠٠) مفحوص، تقل مع زيادة طول الاختبار ونقصان مستوى الدلالة الإحصائية عند جميع مستويات طول الاختبار.

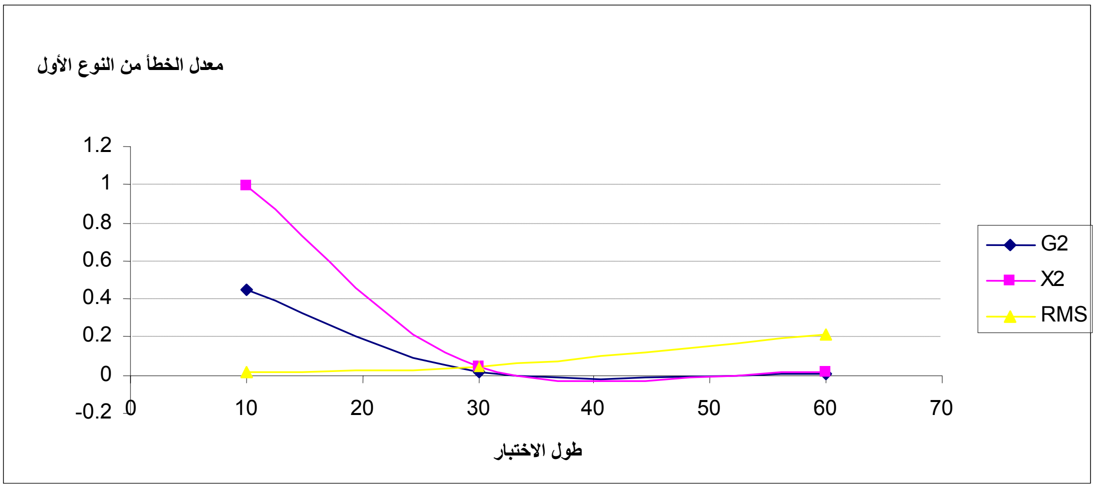


الشكل ٧. الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة (عند حجم عينة ١٠٠فرد) ومستوى دلالة (٠,٠٥) تحت افتراض النموذج اللوجستي ثلاثي المعالم

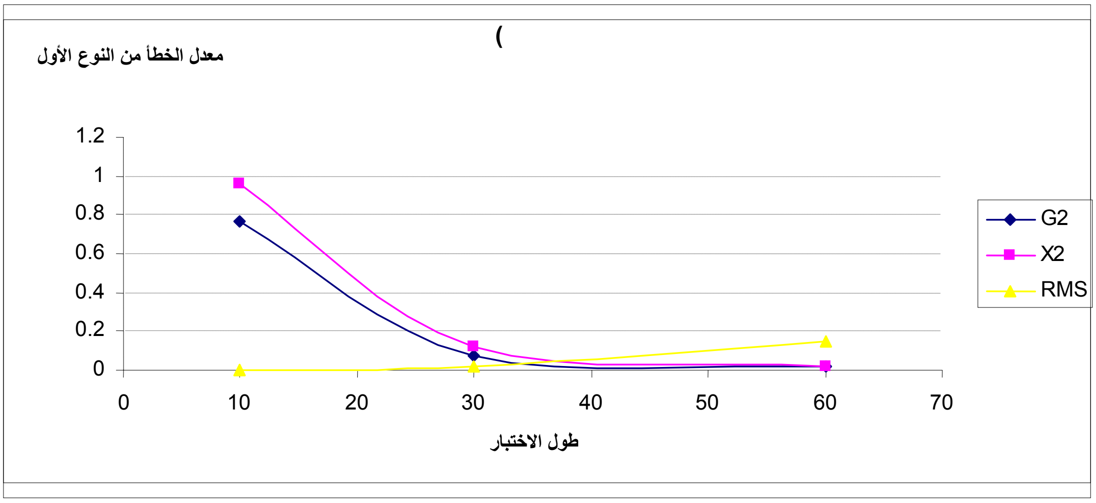
- في حين أن الخطأ من النوع الأول لإحصائي جودة مطابقة (RMS) عند نفس حجم العينـة وعند جميع مستويات الدلالة الإحصائية يزيد مع زيادة طول الاختبار ،وكما يوضحها الشكل (٧).
- نجد بأن الإحصائي (G^2) بشكل عام يتمتع بخطأ منخفضة مقارنة بأداء الإحصائي (χ^2_B)، في حين أبدى إحصائي البواقي المعيارية (RMS) أداء أفضل من الإحـصائيين الآخـرين وبخطأ تصل لصفر عند مستوى دلالة | Z=٣| وطـول اختبـار (١٠) فقـرات ،وكـان إحصائي البواقي المعيارية (RMS) الأسوأ في الأداء عند طول اختبار٣٠ فقرة و٦٠ فقرة مقارنة بالإحصائيين (χ^2_B) و(G^2).

* وعند حجم عينة (٥٠٠) مفحوص و(١٠٠٠) مفحوص كانت نتائج أداء الإحصائيات(G^2) و (χ^2_B) مماثلة لظرف السابق، فنلاحظ أن الخطأ من النوع الأول تقل مـع زيـادة طـول الاختبار ونقصان مستوى الدلالة الإحصائية عند جميع مستويات طول الاختبار، كما هـي

مبينة بالجدول (٤) ،وتوضحها الأشكال (٨) و(٩) في حين يبدي الإحصائي (RMS) سلوكا معاكسا وذلك بزيادة الخطأ من النوع الأول بزيادة طول الاختبار، ونلاحظ أن الإحـصائي (RMS) يتمتع بخطأ منخفضة تصل تقريباً لصفر عند الاختبارات القـصيرة بطـول (١٠) فقرات. وكان أداء الإحصائي (RMS) أفضل من الإحصائيين الآخرين عند هذا الظـرف ،ولكن بزيادة طول الاختبار فإن أداء هذا الإحصائي ينخفض بشكل حاد، بزيادة الخطأ مـن النوع الأول عند زيادة طول الاختبار إلى (٣٠) فقرة و(٦٠) فقرة، في حين يتحسن أداء كل من الإحصائيين (G^2) و (χ^2_B) مع زيادة طول الاختبار ،ويتبين أن الإحـصائي (G^2) يتمتع بخطأ من النوع الأول أقل عند جميع مستويات الدلالة الإحصائية مقارنة بالإحصائيين بالآخرين ويأتي بالمرتبة الثانية إحصائي (χ^2_B).



الشكل ٨. الخطأ من النوع الأول جودة مطابقة الفقرة (عند حجم عينة ٥٠٠فرد) ومستوى دلالة (٠,٠٥) تحت افتراض النموذج اللوجستي ثلاثي المعالم



الشكل ٩. الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة (عند حجم عينة ١٠٠٠فرد) ومستوى دلالة (٠,٠٥) تحت افتراض النموذج اللوجستي ثلاثي المعالم

*** للإجابة على الجزء الذي يتعلق بمعرفة أثر طول الاختبار على قوة الاختبار لإحـصائيات جودة مطابقة الفقرة عند ثبات متغير حجم العينة.**

اتبعت آلية توليد الفقرات المذكورة سابقا لفحص قوة الاختبار ،وحسبت قوة الاختبـار عند كل ظرف، من خلال حساب عدد الحالات غير المطابقة وفق مستويات الدلالـة المحـدد وهي(٠,١ ، ٠,٠٥ ، ٠,٠١)، واعتبرت بأنها فقرات غير مطابقة للنموذج الرياضـي ،وتـم حساب نسبة هذه الحالات غير المطابقة من عدد الحالات المولدة ،مـن خـلال قـسمة عـدد الفقرات غير المطابقة على عدد فقرات الاختبار مضروبا في عدد مـرات الإعـادة فعلـى سبيل المثال :الخلية الأولى في الجدول (٥) عند حجم عينة ١٠٠ ومفحوص وطول اختبـار ١٠ فقرات كانت القوة للإحصائي (G^2)عند مستوى دلالة(٠,١) هـو (٠,٠٨) . ويمكـن تلخيص العملية الحسابية بالخطوات التالية:

١-توليد بيانات لاختبار بطول ١٠ فقرات وتم إعادة التوليد مائة مرة.

٢-حساب قيمة الإحصائي(G^2) ودلالاتها الإحصائية لكل فقرة من الفقرات الناتجة في عـدد المرات المولدة ،حيث بلغ عددها في جميع مرات التوليد (١٠٠٠)فقرة .

٣- مثل توزيع المعاينة لقيم الإحصائي(G^2) وحسبت عدد الحالات التي قلت أو تساوي في قيمتها عن مستوى الدلالة(٠,١) واعتبرت بأنها فقرات غير مطابقة للنموذج،حيث كان عددها (٧٨)فقرة من أصل (١٠٠٠)فقرة .

٤- حسبت النسبة تحت منحى توزيع المعاينة و التي تمثل قوة الاختبار بإيجاد نسبة عدد الفقرات غير المطابقة عبر مائة مرة إعادة مقسوما على مضروب طول الاختبار في عدد مرات الإعادة.وقوة الإحصائي =٧٨/(١٠*١٠٠)=٠,٠٨.

ولمزيد من الإيضاح يمكن تفصيل لقيمة الواردة في الخلية الثانية في الجدول (٥) عند حجم عينة ١٠٠ ومفحوص وطول اختبار ٣٠ فقرة ،كانت القوة للإحصائي (G^2)عند مستوى دلالة(٠,١) هو (0.0٣) .وهذه القيمة كانت ناتج عن الخطوات التالية:

١-توليد بيانات لاختبار بطول ٣٠ فقرة وتم إعادة التوليد مائة مرة.

٢-حساب قيمة الإحصائي(G^2) ودلالاتها الإحصائية لكل فقرة من الفقرات الناتجة في عدد المرات المولدة ،حيث بلغ عددها في جميع مرات التوليد (٣٠٠٠)فقرة .

٣- مثل توزيع المعاينة لقيم الإحصائي(G^2) وحسبت عدد الحالات التي قلت أو تساوي في قيمتها عن مستوى الدلالة(٠,١) واعتبرت بأنها فقرات غير مطابقة للنموذج،حيث كان عددها (٩٨)فقرة من أصل (٣٠٠٠)فقرة .

٤- حسبت النسبة تحت منحى توزيع المعاينة و التي تمثل قوة الاختبار بإيجاد نسبة عدد الفقرات غير المطابقة عبر مائة مرة إعادة مقسوما على مضروب طول الاختبار في عدد مرات الإعادة.وقوة الإحصائي=٩٨/(٣٠*١٠٠)=٠,٠٣.

و الخلية الثالثة في الجدول (٥) عند حجم عينة ١٠٠ ومفحوص وطول اختبار ٦٠ فقرة ،كانت القوة للإحصائي (G^2)عند مستوى دلالة(٠,١)هو (٠,٠٤) .وهذه القيمة كانت ناتج عن الخطوات التالية:

١-توليد بيانات لاختبار بطول ٦٠ فقرة وتم إعادة التوليد مائة مرة.

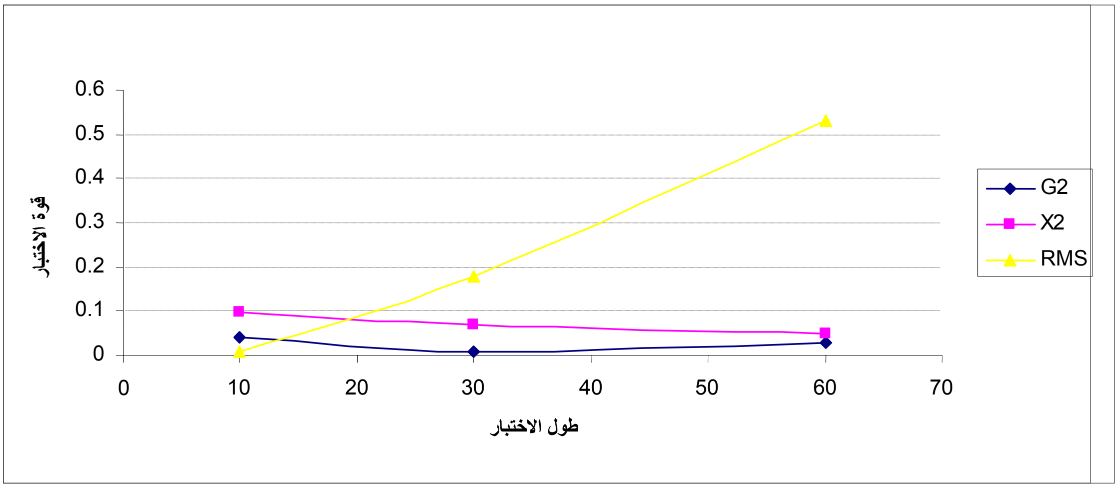
٢-حساب قيمة الإحصائي(G^2) ودلالاتها الإحصائية لكل فقرة من الفقرات الناتجة في عدد المرات المولدة ،حيث بلغ عددها في جميع مرات التوليد (٦٠٠٠)فقرة .

٣- مثل توزيع المعاينة لقيم الإحصائي(G^2) وحسبت عدد الحالات التي قلت أو تساوي في قيمتها عن مستوى الدلالة(٠,١) واعتبرت بأنها فقرات غير مطابقة للنموذج،حيث كان عددها (٢٥٩)فقرة من أصل (٦٠٠٠)فقرة .

٤- حسبت النسبة تحت منحى توزيع المعاينة و التي تمثل قوة الاختبار بإيجاد نـسبة عـدد الفقرات غير المطابقة عبر مائة مرة إعادة مقسوما على مضروب طول الاختبار في عدد مرات الإعادة.وقوة الإحصائي=٢٥٩/(٦٠*١٠٠)=٠,٠٤

الجدول ٥. قوة الاختبار لإحصائيات جودة مطابقة الفقرة

حجم العينة	طول الاختبار (عدد الفقرات)	إحصائي جودة مطابقة الفقرة								
		إحصائي (G^2) الأرجحية العظمى.			إحصائي Bocks (χ^2_B)			وإحصائي البواقي المعيارية (RMS)		
		مستوى الدلالة(α)			مستوى الدلالة(α)			مستوى الدلالة(α)		
		٠,١	٠,٠٥	٠,٠١	٠,١	٠,٠٥	٠,٠١	Z=١	Z=٢	Z=٣
١٠٠	١٠	٧٨ (٠,٠٨)	٤٣ (٠,٠٤)	٢٤ (٠,٠٢)	١٣٣ (٠,١٣٣)	٩٨ (٠,٠٩٨)	٥٥ (٠,٠٥٥)	٣٢٠ (٠,٣٢)	٩ (٠,٠٠٩)	صفر (صفر)
	٣٠	٩٨ (٠,٠٣)	٤١ (٠,٠١)	١٥ (٠,٠٠٥)	٤١٥ (٠,١٤)	٢١٠ (٠,٠٧)	٤٩ (٠,٠٢)	٢١٤٨ (٠,٧٢)	٥٥٢ (٠,١٨)	١٠٨ (٠,٤)
	٦٠	٢٥٩ (٠,٠٤)	١٧٦ (٠,٠٣)	١٠٣ (٠,٠٢)	٥٦٨ (٠,٠٩٥)	٢٨٤ (٠,٠٥)	٤٢ (٠,٠٠٧)	٥٧٨٠ (٠,٩٦)	٣١٦٨ (٠,٥٣)	٩٢٢ (٠,١٥)
٥٠٠	١٠	٣٦١ (٠,٣٦)	٢٣٧ (٠,٢٣)	٩٠ (٠,٠٩)	٩٨٤ (٠,٩٨)	٩٨٣ (٠,٩٨)	٩٧٦ (٠,٩٨)	٣٤٦ (٠,٣٥)	٦ (٠.٠٠٦)	صفر (صفر)
	٣٠	٤٥١ (٠,١٥)	٢٨٣ (٠,٠٩)	١١٠ (٠,٠٤)	٧٧٢ (٠,٢٦)	٥٢٢ (٠,١٧)	٢٠٧ (٠,٠٧)	١٩٧٣ (٠,٦٦)	٥٠٣ (٠,١٧)	٥٩ (٠,١٧)
	٦٠	٤١٦ (٠,٠٧)	٢٥١ (٠,٠٤)	٤٨ (٠,٠٠٨)	٧٢٨ (٠,١٢)	٤٠٩ (٠,٠٧)	١٠٤ (٠,٠٢)	٥٤٨٤ (٠,٩١)	٢٠٨٨ (٠,٣٥)	٣٩٣ (٠,٠٧)
١٠٠٠	١٠	٧٧٥ (٠,٧٨)	٦٥٨ (٠,٦٦)	٤٣٥ (٠,٤٤)	١٠٠٠ (١)	١٠٠٠ (١)	١٠٠٠ (١)	٦٦٣ (٠,٦٦)	٦٩ (٠,٠٧)	صفر (صفر)
	٣٠	٦٢٢ (٠,٢١)	٣٨٨ (٠,١٣)	١٠٦ (٠,٠٤)	٩٧٥ (٠,٣٣)	٦٨٨ (٠,٢٣)	٢٦٦ (٠,٠٨٩)	٢٣٣١ (٠,٧٨)	١٠٠٦ (٠,٣٤)	٢٥٠ (٠,٠٨٣)
	٦٠	٨٠٧ (٠,١٤)	٤٧٦ (٠,٠٨)	١٣٠ (٠,٠٢)	١١٨٨ (٠,١٩٨)	٣٢٩ (٠,٠٦)	٢٧٠ (٠,٠٤٥)	٥٥٠٧ (٠,٠٩٢)	٢٦٠٣ (٠,٤٣)	٨٣٦ (٠,١٤)



الشكل ١٠. قوة الاختبار لإحصائيات جودة مطابقة الفقرة (عند حجم عينة ١٠٠فرد) ومستوى دلالة (٠,٠٥)

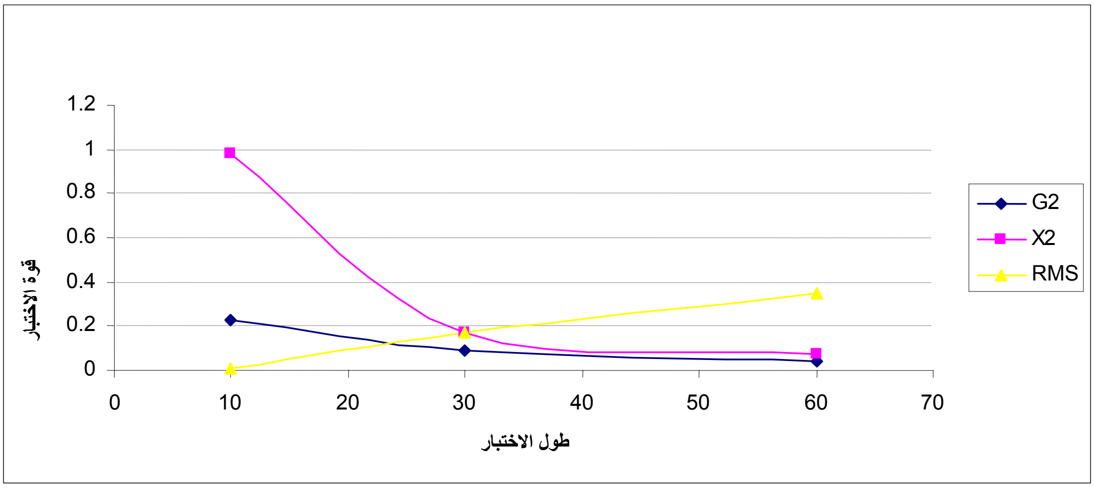
يبين الجدول (٥) قوة الاختبار لإحصائيات جودة مطابقة الفقرة وهي إحصائي الأرجحية العظمى (G^2-index) وإحصائي بك (Bock-index) (χ^2_B) وإحصائي البواقي المعيارية (RMS)، وتمثل القيم في الجدول المساحة تحت منحى توزيع المعاينة و التي تعكس التكرارات والنسبة المئوية لعدد الفقرات غير المطابقة عبر مائة مرة إعادة، حيث تم توليد الاستجابات على الفقرات ومعايرتها(Calibration) تحت افتراض النموذج النموذج ثلاثي المعالم ومعايرتها(Calibration) بنموذج مختلف هو نموذج ثنائي المعالم.

ويبين الجدول (٥) :

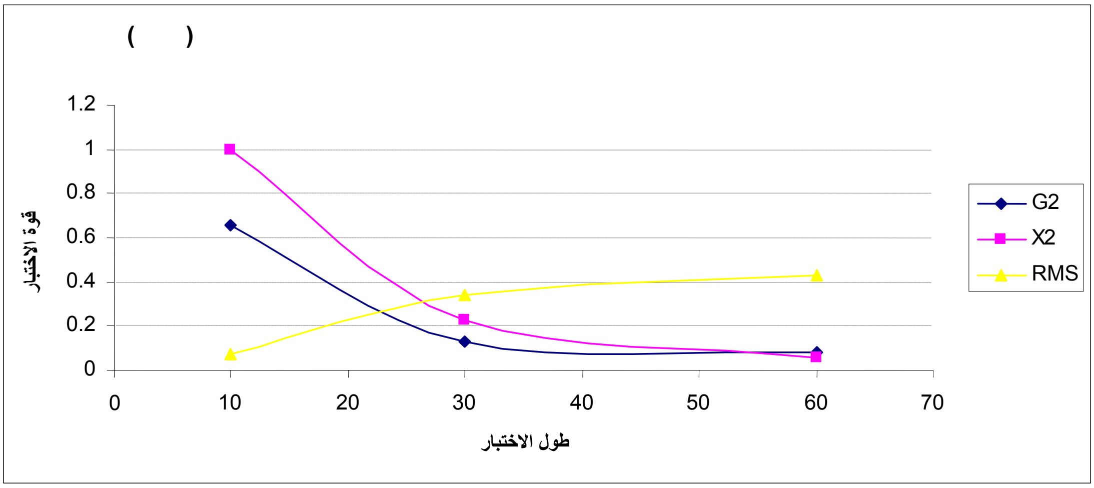
* أن قوة الاختبار لإحصائيات جودة مطابقة الفقرة (G^2) و (χ^2_B) عند حجم عينة (١٠٠) مفحوص، تقل مع زيادة طول الاختبار ونقصان مستوى الدلالة الإحصائية عند جميع مستويات طول الاختبار.

- في حين أن قوة الاختبار لإحصائي جودة مطابقة (RMS) عند نفس حجم العينة وعند جميع مستويات الدلالة الإحصائية يزيد مع زيادة طول الاختبار.
- ونلاحظ بأن الإحصائي (χ^2_B) يتمتع بقوة إحصائية أكبر مقارنة بأداء الإحصائي (G^2)، عند طول اختبار٣٠ فقرة و٦٠ فقرة في حين أبدى إحصائي البواقي المعيارية (RMS) قوة أفضل عند مستوى دلالة | Z=١| وطول اختبار (٦٠) فقرة وبشكل عام فإن القوة الإحصائية للإحصائيين (χ^2_B) و(G^2) منخفضة عند هذا الظرف والشكل (١٠) يوضح ذلك.

وعنـد حجـم عينـة (٥٠٠) مفحـوص و(١٠٠٠) مفحـوص كانـت نتـائج أداء الإحصائيات(χ^2_B) و(G^2) مماثلة للظرف السابق، إذ نلاحظ أن القوة الإحصائية لإحـصائيات جودة مطابقة الفقرة تقل مع زيادة طول الاختبار ونقصان مستوى الدلالة الإحصائية عند جميع مستويات طول الاختبار، في حين يبدي الإحصائي (RMS) سلوكاً معاكساً وذلك بزيادة القوة الإحصائية بزيادة طول الاختبار، وكانت قوة الإحصائي (RMS) أكبر من الإحصائيين (χ^2_B) و(G^2) على الاختبارات الطويلة (٣٠، ٦٠) فقرة ، في حين أن القوة الإحصائية للإحـصائي (χ^2_B) تصل إلى (١) عند الاختبارات القصيرة (١٠) فقرات. وأداء الإحصائي (χ^2_B) أفـضل من أداء (G^2) بشكل عام عند جميع مستويات طـول لاختبـار كمـا يبينهـا الجـدول (٥) وتوضحها الأشكال (١١) (١٢).



الشكل ١١. قوة الاختبار لإحصائيات جودة مطابقة الفقرة (عند حجم عينة ٥٠٠فرد) ومستوى دلالة (٠,٠٥)



الشكل ١٢. قوة الاختبار لإحصائيات جودة مطابقة الفقرة (عند حجم عينة ١٠٠٠فرد) ومستوى دلالة (٠,٠٥)

- النتائج المتعلقة بالسؤال الثاني" هل تختلف الخطـأ مـن النـوع الأول وقـوة الاختبـار الإحصائي لإحصائيات لفحص جودة مطابقة الفقرة باختلاف حجم العينة ؟."

* للإجابة على الجزء الذي يتعلق بمعرفة أثر حجم العينة علـى الخطـأ مـن النـوع الأول لإحصائيات جودة مطابقة الفقرة عند ثبات طول الاختبار.

يبين الجدول (٦) الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة، عنـد كـل خلية تمثلت بتفاعل طول الاختبار مع حجم العينة ،وتمثل القيم في الجـدول المـساحة تحـت منحى توزيع المعاينة و التي تعكس التكرارات والنسبة المئوية لعدد الفقرات غيـر المطابقـة عبر مائة مرة إعادة، حيث تمت توليد الاستجابات على الفقـرات ومعايرتهـا(Calibration) تحت افتراض النموذج ثنائي المعالم.

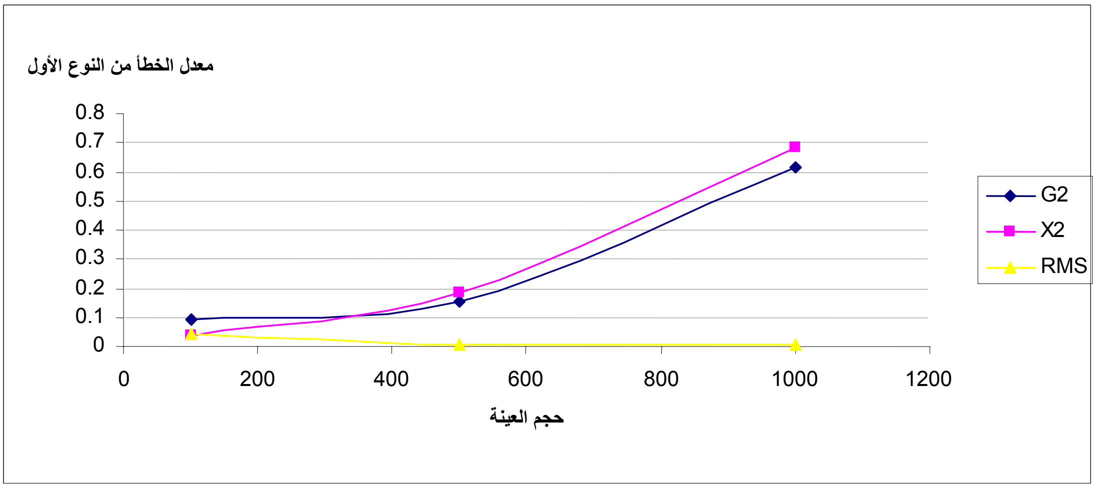
الجدول ٦. الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة تحت افتراض النموذج اللوجستي ثنائي المعالم تبعاً لاختلاف حجم العينة

طول الاختبار (عدد الفقرات)	حجم العينة	إحصائي جودة مطابقة الفقرة								
		إحصائي (G^2) الأرجحية العظمى.			إحصائي Bocks (χ^2_B)			إحصائي البواقي المعيارية (RMS)		
		مستوى الدلالة(α)			مستوى الدلالة(α)			مستوى الدلالة Z		
		٠,١	٠,٠٥	٠,٠١	٠,١	٠,٠٥	٠,٠١	Z=١	Z=٢	Z=٣
١٠	١٠٠	١٣٨ (٠,١٣٨)	٩٠ (٠,٠٩)	٥٨ (٠,٠٥٨)	٧٥ (٠,٠٧٥)	٣٨ (٠,٠٣٨)	١٣ (٠,٠١٣)	٥٠٣ (٠,٥٠٣)	٤٦ (٠,٠٤٦)	١١ (٠,٠١١)
	٥٠٠	٢١٩ (٠,٢١٩)	١٥١ (٠,١٥١)	٦٦ (٠,٠٦٦)	٢٨١ (٠,٢٨١)	١٨٢ (٠,١٨٢)	٦٧ (٠,٠٦٧)	٢٢٥ (٠,٢٢٥)	٧ (٠,٠٠٧)	صفر (صفر)
	١٠٠٠	٧٧٤ (٠,٧٧٤)	٦١٣ (٠,٦١٣)	٣٢٣ (٠,٣٢٣)	٨٠٧ (٠,٨٠٧)	٦٨٥ (٠,٦٨٥)	٤٤٣ (٠,٤٤٣)	٩٦ (٠,٠٩٦)	٤ (٠,٠٠٤)	صفر (صفر)
٣٠	١٠٠	٢٣٩ (٠,٠٧٩)	٢١٦ (٠,٠٧٢)	١٧٨ (٠,٠٥٩)	٩٠ (٠,٠٣)	٤١ (٠,٠٤١)	٥ (٠,٠٠٢)	١٦١٢ (٠,٥٣٧)	٤٩٥ (٠,١٦٥)	٢٠٨ (٠,٠٦٩)
	٥٠٠	١١٤ (٠,٠٣٨)	٥٥ (٠,٠١٨)	١٤ (٠,٠٠٥)	٤٠٣ (٠,١٣٤)	٢١٦ (٠,٠٧٢)	٥٣ (٠,٠١٨)	٢٤٥٥ (٠,٨١٨)	٨١٣ (٠,٢٧)	١٧٤ (٠,٠٥٨)
	١٠٠٠	١٨٣ (٠,٦١)	٨٩ (٠,٢٩)	٩ (٠,٠٠٣)	٦٣٨ (٠,٢١٣)	٣٧٣ (٠,١٢٤)	٩٥ (٠,٠٣٢)	٢٣٧٤ (٠,٧٩)	٦٢٥ (٠,٢١)	٧٥ (٠,٠٣)
٦٠	١٠٠	٩٨ (٠,٠١٦)	٣٥ (٠,٠٠٦)	٥ (٠,٠٠٠٨)	١٥٤ (٠,٠٢٦)	٧٠ (٠,٠١٢)	١٢ (٠,٠٠٢)	٥٨٩١ (٠,٩٨٢)	٤٣٨٩ (٠,٧٣٢)	٢٠٨٧ (٠,٣٤٨)
	٥٠٠	١٧٢ (٠,٠٢٩)	٧٣ (٠,٠١٢)	١١ (٠,٠٠٢)	٣٣٩ (٠,٠٥٧)	١٦٣ (٠,٠٢٧)	٢٧ (٠,٠٠٥)	٥٩٣٠ (٠,٩٩)	٣٧٩٠ (٠,٦٣)	١٢٦٤ (٠,٢١)
	١٠٠٠	٢٣٨ (٠,٠٣٩)	١٠٨ (٠,٠١٨)	١٧ (٠,٠٠٣)	٣٣٨ (٠,٠٥٦)	١٥٦ (٠,٠٢٦)	٢٦ (٠,٠٠٤)	٥٩٢٩ (٠,٠٠٤)	٣٢٦٠ (٠,٥٤)	٦٢٣ (٠,١٠٤)

ويتضح من الجدول (٦):

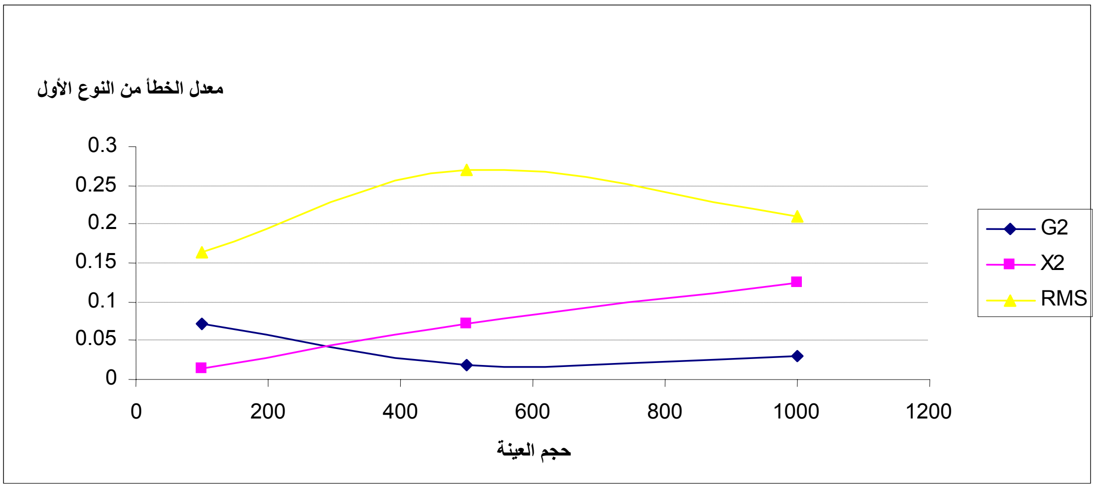
* أن الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة (G^2) و (χ^2_B) عند طول اختبار (١٠) فقرات، تزيد مع زيادة حجم العينة وزيادة مستوى الدلالة الإحصائية عند جميع مستويات حجم العينة.

* في حين أن الخطأ من النوع الأول لإحصائي جودة مطابقة البواقي المعيارية (RMS) عند نفس طول الاختبار وعند جميع مستويات الدلالة الإحصائية يقل مع زيادة حجم العينة.
* ونجد بأن الإحصائي بك (χ^2_B) يتمتع بخطأ من النوع الأول أقل من الإحـصائي (G^2)، عند حجم عينة (١٠٠) مفحوص، في حين أبدى الإحصائي (G^2) معـدلات خطـأ أقـل مقارنة بالإحصائي (χ^2_B) عند حجم عينة (٥٠٠) (١٠٠٠) مفحوص. في حـين أظهـر الإحصائي (البواقي المعيارية) (RMS)أداء أفضل مـن الإحـصائيين (χ^2_B) و(G^2)، إذ تميز بخطأ منخفضة تصل لصفر عند مستويات دلالة | Z=٣| ويأتي الإحصائي(RMS) بالمرتبة الأولى إذ جاء بخطأ منخفضة عنـد مـستويات دلالـة | Z=٢| و | Z=٣| مقارنة بالإحصائيين الآخرين وكما بينها الشكل (١٣).



الشكل ١٣. الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة (عند طول اختبار ١٠ فقرة) ومستوى دلالة (٠,٠٥) تحت افتراض النموذج اللوجستي ثنائي المعالم

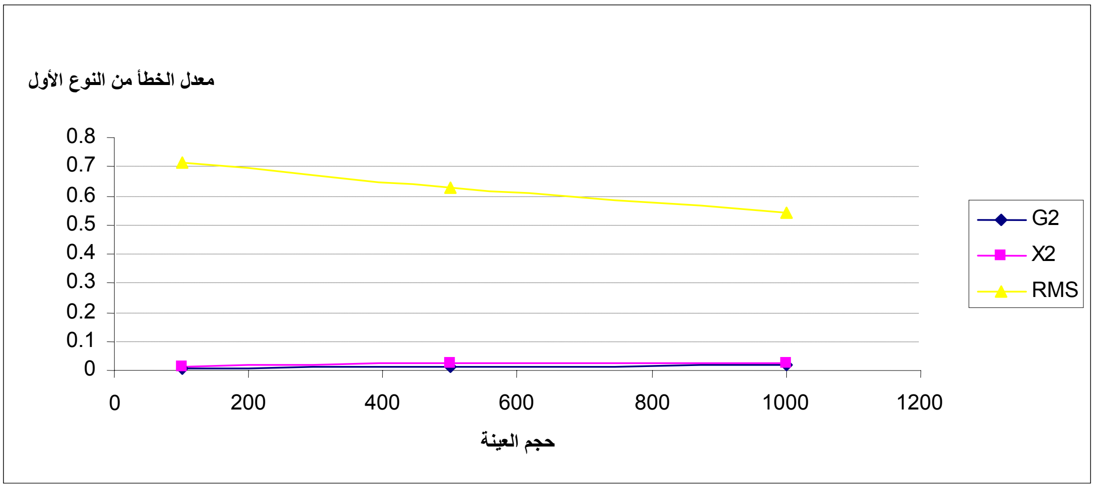
*وعند طول اختبار (٣٠) فقرة، نلاحظ أن الخطأ من النوع الأول للإحصائي (χ^2_B) تزداد مع زيادة حجم العينة، في حين أن الإحصائي (G^2) تنخفض الخطأ من النـوع الأول عنـد حجم عينة (٥٠٠) و(١٠٠٠) مفحوص، ولكنها تعاود الارتفاع عند حجم عينة (١٠٠) مفحوص، في حين يبدي الإحصائي (RMS) سلوكاً معاكساً عند مـستوى دلالـة | Z=٣|، وذلـك بنقصان الخطأ من النوع الأول بزيادة حجم العينة فيما يتشابه مع إحـصائي (G^2) عنـد مستوى دلالة | Z=١| و | Z=٢| كما يتضح في الشكل (١٤).



الشكل ١٤. الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة (عند طول اختبار ٣٠ فقرة) ومستوى دلالة (٠,٠٥) تحت افتراض النموذج اللوجستي ثنائي المعالم

* وعند طول اختبار (٦٠) فقرة، نلاحظ أن الخطأ من النوع الأول للإحصائيين (G^2) (χ^2_B) تميزت بخطأ منخفضة عند جميع مستويات حجم العينة، وكان أفضلها عنـد حجـم عينـة (١٠٠) مفحوص فيما كانت متقاربة عند حجم عينة (٥٠٠ و ١٠٠٠) مفحوص مع ملاحظة زيادة الخطأ بزيادة حجم العينة.

* أما فيما يتعلق بإحصائي جودة المطابقة (RMS) تميز بخطأ من النوع الأول مرتفعة عند جميع مستويات حجم العينة وكان سلوكه معاكساً للإحصائيات الأخرى فنجد أن الخطأ مـن النوع الأول يقل مع زيادة حجم العينة وكما يوضحها الشكل (١٥).



الشكل ١٥. الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة (عند طول اختبار ٦٠ فقرة) ومستوى دلالة (٠,٠٥) تحت افتراض النموذج اللوجستي ثنائي المعالم

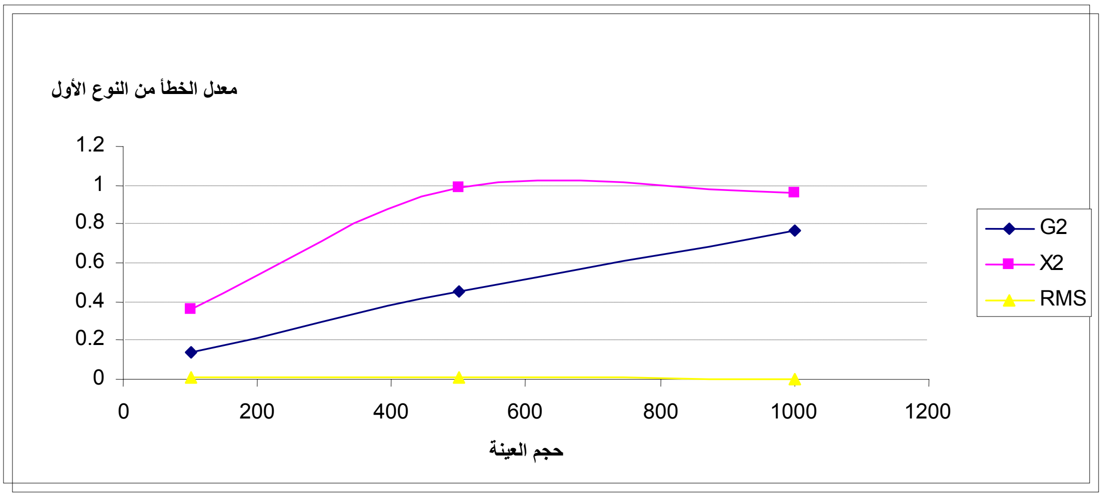
ويبين الجدول (٧) :الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقـرة، وهـي إحصائي الأرجحية العظمى (G^2-index) وإحصائي بك (Bock-index) (χ^2_B) وإحـصائي البواقي المعيارية (RMS) ،وتمثل القيم في الجدول المساحة تحت منحى توزيـع المعاينـة و التي تعكس التكرارات والنسبة المئوية لعدد الفقرات غير المطابقة عبر مائة مرة إعادة، حيث تمت توليد الاستجابات على الفقرات ومعايرتها(Calibration)تحت افتراض النموذج ثلاثـي المعالم

الجدول ٧. الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة تحت افتراض النموذج اللوجستي ثلاثي المعالم

طول الاختبار (عدد الفقرات)	حجم العينة	إحصائي جودة مطابقة الفقرة								
		إحصائي (G^2) الأرجحية العظمى.			إحصائي Bocks (χ^2_B)			وإحصائي البواقي المعيارية (RMS)		
		مستوى الدلالة(α)			مستوى الدلالة(α)			مستوى الدلالة(α)		
		٠,١	٠,٠٥	٠,٠١	٠,١	٠,٠٥	٠,٠١	١=Z	٢=Z	٣=Z
١٠	١٠٠	١٨٢ (٠,١٨٣)	١٤٣ (٠,١٤٣)	٩٦ (٠,٠٩٦)	٤٩٤ (٠,٤٩٤)	٣٦٤ (٠,٣٦)	١٥٧ (٠,١٥٧)	١٨٠ (٠,١٨)	٧ (٠,٠٠٧)	صفر (صفر)
	٥٠٠	٥٥٥ (٠,٥٦)	٤٥١ (٠,٤٥)	٣٠٥ (٠,٣)	٩٩٠ (٠,٩٩)	٩٨٨ (٠,٩٩)	٩٨٦ (٠,٩٩)	٢٣٣ (٠,٢٣٣)	١٢ (٠,٠١٢)	٢ (٠,٠٠٢)
	١٠٠٠	٨٥٩ (٠,٨٦)	٧٦٨ (٠,٧٧)	٥٦٩ (٠,٥٧)	٩٦٣ (٠,٩٦)	٩٦٣ (٠,٩٦)	٩٦٢ (٠,٩٦)	٥٣ (٠,٠٥٣)	صفر (صفر)	صفر (صفر)
٣٠	١٠٠	١٩٧ (٠,٠٦٦)	٨٥ (٠,٠٣)	٤ (٠,٠٠١)	١٠٢ (٠,٠٣)	٣٩ (٠,٠١)	١ (٠,٠٠٠٣)	٢٠٦٤ (٠,٦٩)	٤٥٨ (٠,١٥)	١١٤ (٠,٠٣٨)
	٥٠٠	١٤٤ (٠,٠٥)	٦٣ (٠,٠٢)	١٠ (٠,٠٠٣)	٢٦٠ (٠,٠٩)	١٢٧ (٠,٠٤)	١٩ (٠,٠٠٦)	١٤٤٦ (٠,٤٨)	١٢٣ (٠,٠٤١)	٢٣ (٠,٠٠٨)
	١٠٠٠	٣٨١ (٠,١٣)	٢٣٣ (٠,٠٧٨)	٨١ (٠,٠٣)	٥٥٠ (٠,١٨)	٣٤٩ (٠,١٢)	١٠٣ (٠,٠٣)	٢٢٥٣ (٠,٧٥)	٧٢ (٠,٠٢)	٧ (٠,٠٠٢)
٦٠	١٠٠	١٤٧ (٠,٠٣)	٧٩ (٠,٠١)	٤٥ (٠,٠٠٨)	١٩٨ (٠,٠٣)	٨٥ (٠,٠١)	٥ (٠,٠٠٠٨)	٥٨٥٦ (٠,٩٨)	٣١٩٣ (٠,٥٣)	٩٩٨ (٠,١٦)
	٥٠٠	١٩٠ (٠,٠٣)	٧٨ (٠,٠١)	١٠ (٠,٠٠٢)	٢٩٣ (٠,٠٥)	١٢٩ (٠,٠٢)	١٩ (٠,٠٠٣)	٥٤٣٣ (٠,٩١)	١٢٥٠ (٠,٢١)	١١٧ (٠,٠٢)
	١٠٠٠	٢٤٦ (٠,٠٤)	٩٣ (٠,٠٢)	١٤ (٠,٠٠٢)	٢٩٧ ٠,٠٥	١٣٧ (٠,٠٢)	٢٦ (٠,٠٠٤)	٥٢٤٦ (٠,٨٧)	٨٨٣ (٠,١٥)	٤٤ (٠,٠٠٧)

*ويتضح من الجدول أن الإحصائي (G^2) يتميز بخطأ من النـوع الأول منخفـضة مقارنـة بالإحصائي (χ^2_B) عند طول اختبار (١٠) فقرات، وحجم عينة (١٠٠) مفحوص.

* في حين أبدى الإحصائي (RMS) الخطأ من النوع الأول منخفضة مقارنـة بالإحـصائيين (G^2 ، χ^2_B) لتصل إلى صفر ، عند حجم عينة (١٠٠٠) مفحوص عند الدلالة الإحـصائية | Z=٢| و | Z=٣| وعند حجم عينة (١٠٠) مفحوص مستوى دلالة | Z=٣| . ويبـدي الإحصائيين (G^2) و (χ^2_B) زيادة في الخطأ من النوع الأول بزيادة حجم العينة في حين أن الإحصائي (RMS) أبدى تناقصاً ملحوظاً في الخطأ من النوع الأول بزيادة حجم العينة من (١٠٠) مفحوص إلى (١٠٠٠) مفحوص في حين عادت الخطأ بالارتفاع عند حجـم عينـة (٥٠٠) مفحوص كما يوضحها الشكل (١٦).

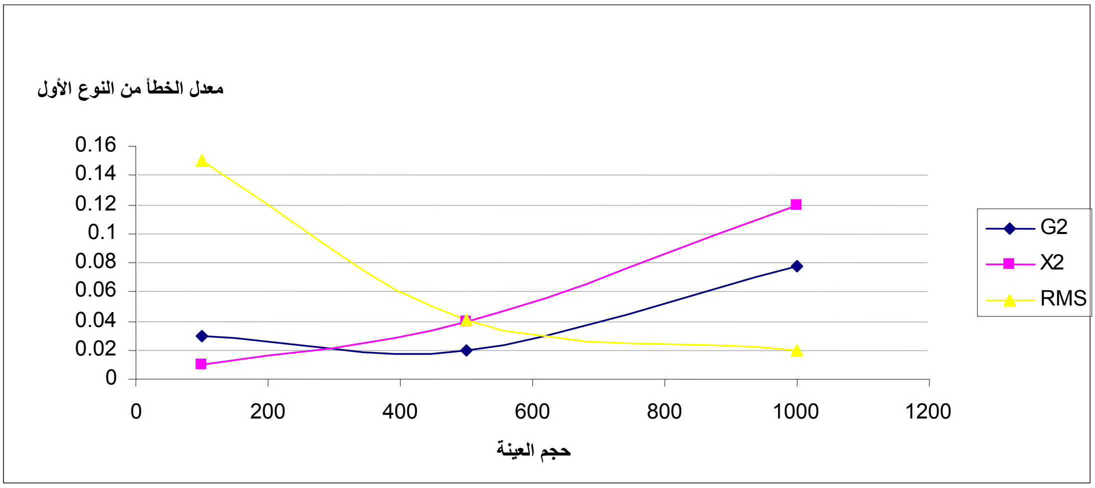


الشكل ١٦. الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة (عند طول اختبار ١٠ فقرة) ومستوى دلالة (٠,٠٥) تحت افتراض النموذج اللوجستي ثلاثي المعالم

*وعند دراسة اداء الإحصائيات عند طول اختبار (٣٠) فقرة، تبين أن الخطأ من النوع الأول للإحصائيات (G^2) (χ^2_B) تزداد مع زيادة حجم العينة، وتميزت بشكل عام بخطأ منخفضة عند جميع مستويات الدلالة.

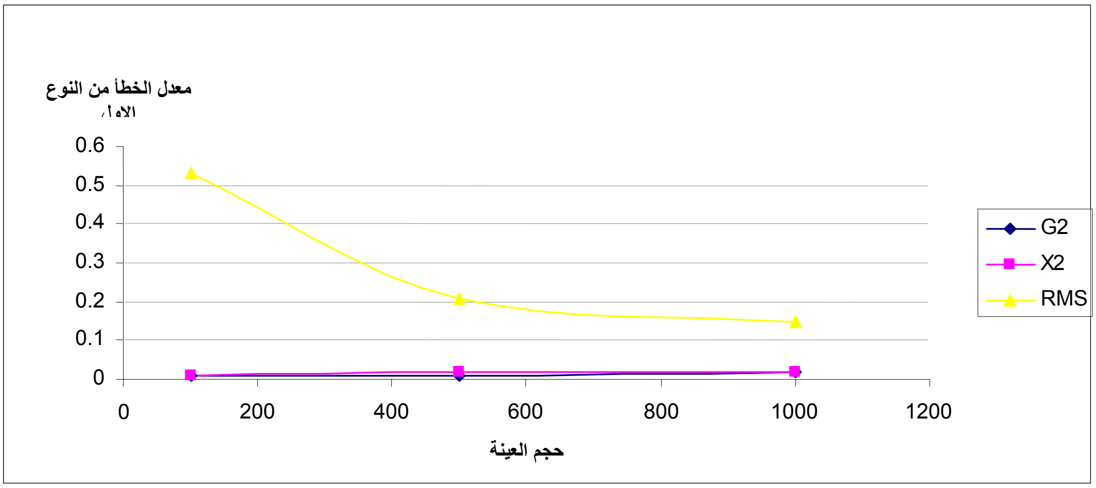
*في حين أن الإحصائي (RMS) تميز بخطأ مرتفعة عند مـستوى دلالـة | Z=١| وعنـد مستوى دلالة | Z=٢| و | Z=٣| أبدى سلوكاً معاكساً بانخفاض الخطأ من النوع الأول مـع

زيادة حجم العينة، وتميزت بخطأ المنخفضة مقارنة مع الإحصائيين (χ^2_B) (G^2) كما يبينها الشكل (١٧).



الشكل ١٧. الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة (عند طول اختبار ٣٠ فقرة) ومستوى دلالة (٠,٠٥) تحت افتراض النموذج اللوجستي ثلاثي المعالم

* وعند طول اختبار (٦٠) فقرة، نلاحظ أن الخطأ من النوع الأول للإحصائي جودة مطابقة الفقرة (G^2) (χ^2_B) تميزت بخطأ منخفضة عند جميع مستويات حجم العينة، وتماثل مستويات الخطأ عند جميع مستويات حجم العينة بحيث كانت الخطأ متقاربة رغم الزيادة في حجم العينة.
* وتميز الإحصائي (RMS) بخطأ من النوع الأول مرتفعة مع ملاحظة تناقص في الخطأ بزيادة حجم العينة عند مستوى دلالة |Z=١| وعند زيادة مستوى الدلالة إلى |Z=٢| تنخفض الخطأ لتصبح في أقل مستوياتها عند حجم عينة (٥٠٠) و(١٠٠٠) وعند مستوى دلالة |Z=٣| كما يوضحها الشكل (١٨).

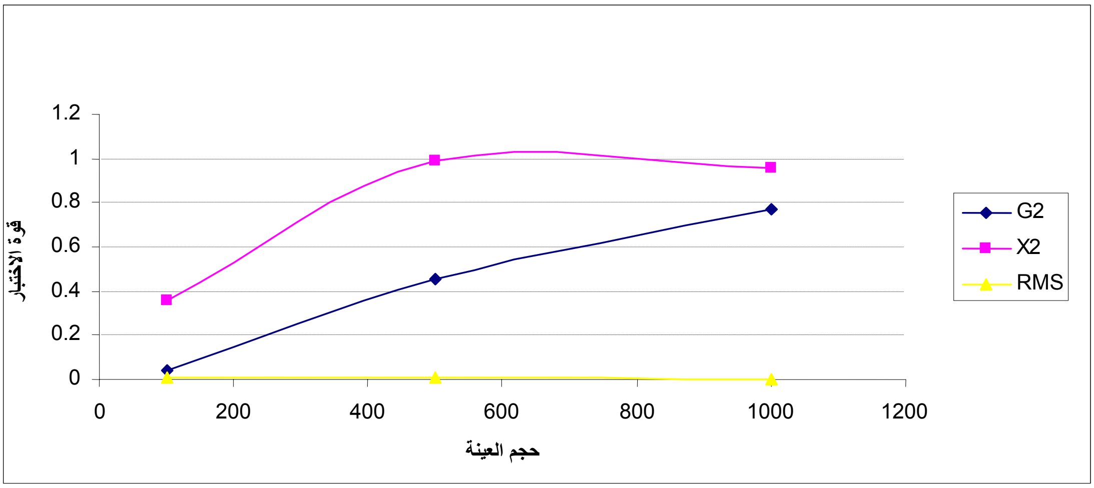


الشكل ١٨. الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة (عند طول اختبار ٦٠ فقرة) ومستوى دلالة (٠,٠٥) تحت افتراض النموذج اللوجستي ثلاثي المعالم

* للإجابة على الجزء الذي يتعلق بمعرفة أثر حجم العينة على قوة الاختبار لإحصائيات جودة مطابقة الفقرة عند ثبات متغير طول الاختبار.

يبين الجدول (٥) : قوة الاختبار لإحصائيات جودة مطابقة الفقـرة وهـي إحـصائي الأرجحية العظمى (G^2) وإحصائي (χ^2_B) وإحصائي البواقي المعيارية (RMS)، وتمثل القيم في الجدول المساحة تحت منحى توزيع المعاينة و التي تعكس التكرارات والنسبة المئوية لعدد الفقرات غير المطابقة عبر مائة مرة إعادة، حيث تمت توليد الاستجابات على الفقرات تحـت افتراض النموذج ثلاثي المعالم ومعايرتها(Calibration) بنموذج مختلف هو نمـوذج ثنـائي المعالم.

* ويلاحظ من الجدول (٥) ويوضحها الشكل (١٩) أن قوة الاختبار لإحصائيات جودة مطابقة الفقرة (χ^2_B) و(G^2) عند طول اختبار عشرة فقرات، تزيد مع زيادة حجم العينة، وتميزت معدلات قوة الاختبار لإحصائي المطابقة (χ^2_B) عند هذا الظرف بالارتفاع الكبير مقارنة مع الإحصائي (G^2) حيث بلغت عند حجم عينة (٥٠٠) و(١٠٠٠) مفحوص تقريباً (١).

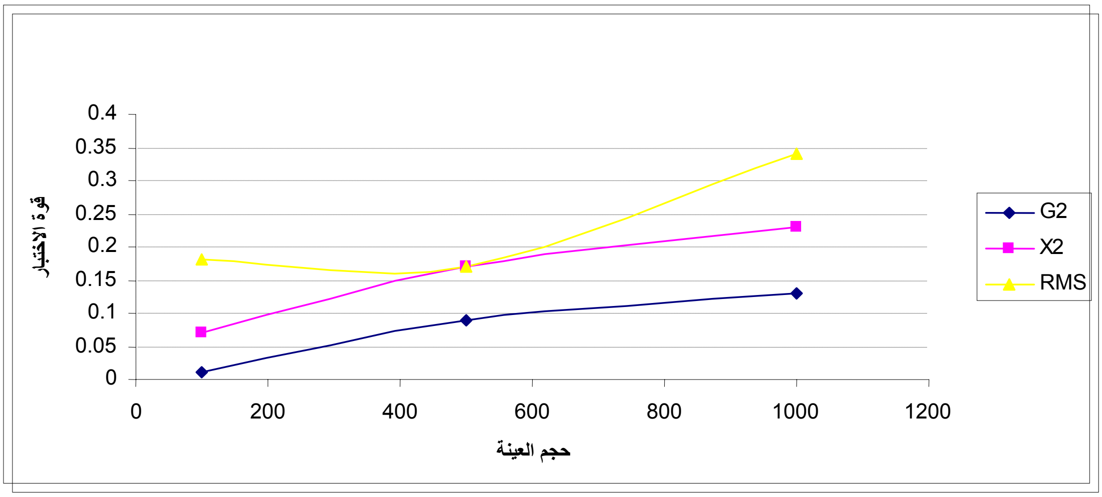


الشكل ١٩. قوة الاختبار لإحصائيات جودة مطابقة الفقرة (عند طول اختبار ١٠ فقرة) ومستوى دلالة (٠,٠٥)

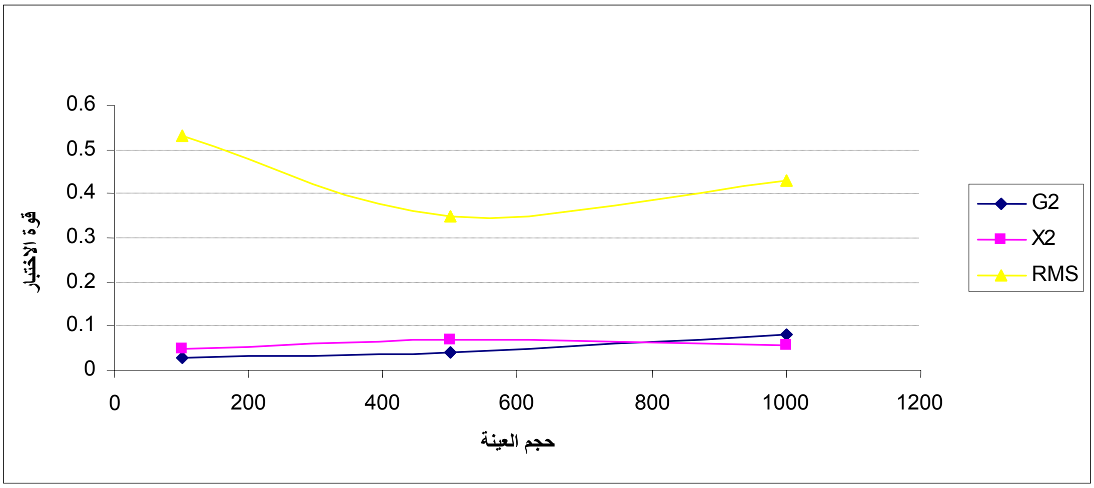
* في حين أن قوة الإحصائي (RMS) زادت بزيادة حجم العينة وتميـزت بالارتفـاع عنـد مستوى دلالة | Z=١| ومع زيادة مستوى الدلالة، أبدى أداء ضعيفاً (مقارنة بالإحـصائيات الأخرى) حيث وصلت القوة للإحصائي (صفر) عند مستوى دلالة | Z=٣| .

* وعند طول اختبار (٣٠) فقرة و(٦٠) فقرة كانت النتائج متقاربة في أداء الإحـصائيات، إذ نلاحظ أن القوة الإحصائية لإحصائيات جودة مطابقة (G^2 ، χ^2_B) تزداد مع زيـادة حجـم العينة ولكنها كانت منخفضة مقارنة عند طول اختبار عشرة فقرات.

* في حين أن الإحصائي (RMS) عند طول اختبار ٣٠ فقرة أبدى سلوكاً معاكـساً وذلـك بزيادة القوة الإحصائية بزيادة طول الاختبار وتميز بمعدلات قوة مرتفعة عند مستوى دلالة | Z=١| في حين تميز بمعدلات قوة منخفضة عند زيادة مستوى الدلالة إلى ٢ أو ٣. وعند طول اختبار ٦٠ فقرة زاد قوة الاختبار بزيادة حجم العينة حيث وصلت إلى (٠,٩١) عنـد حجم عينة (١٠٠٠) بمستوى دلالة | Z=١| و(٠,٤٣) عند مستوى دلالة | Z=٣| وكمـا توضحها الأشكال (٢٠) و(٢١).



الشكل ٢٠. قوة الاختبار لإحصائيات جودة مطابقة الفقرة (عند طول اختبار ٣٠ فقرة) ومستوى دلالة (٠,٠٥)



الشكل ٢١. قوة الاختبار لإحصائيات جودة مطابقة الفقرة (عند طول اختبار ٦٠ فقرة) ومستوى دلالة (٠,٠٥)

- النتائج المتعلقة بالسؤال الثالث" هل تختلف الخطأ من النـوع الأول لإحـصائيات فحـص جودة مطابقة الفقرة باختلاف النموذج الرياضي المستخدم ؟".

يتبين من الجدول (٣) والجدول (٤): الخطأ من النوع الأول تحت افتراض النمـوذج ثنائي المعلم وثلاثي المعلم على التوالي للجدولين لإحصائيات جودة مطابقة الفقرة.

تبين عند مقارنة الخطأ من النوع الأول لإحصائي جودة مطابقة الفقـرة (G^2) فـي الجدولين:

* أن الخطأ من النوع الأول للإحصائي (G^2) تحت افتراض النموذج ثلاثي المعـالم أعلـى بقليل من الخطأ تحت افتراض النموذج ثنائي المعالم وعند جميع مستويات الدلالة، وكـذلك الحال بالنسبة للإحصائي (χ^2_B) .

*في حين نجد أن الإحصائي البواقي المعيارية (RMS) يبدي سلوكاً معاكساً عند معظم الظروف الناتجة عند جميع الخلايا ، فنجد أن الخطأ من النوع الأول كانت أعلى تحت افتراضات النموذج ثنائي المعلم. عدا عند طول اختبار (٣٠) فبدت الخطأ من النوع الأول مرتفعة في النموذج ثلاثي المعالم عنها في النموذج ثنائي المعلم.

***- النتائج المتعلقة بالسؤال الرابع "هل تختلف الخطأ من النوع الأول لإحصائيات فحص جودة مطابقة الفقرة باختلاف التفاعل بين متغيرات: طول الاختبار وحجم العينة؟"**

يتبين من الجدول (٣) والجدول (٤) نتائج الخطأ من النوع الأول تحت افتراض النموذج ثنائي المعلم والنموذج ثلاثي المعالم على التوالي للجدولين لإحصائيات جودة مطابقة الفقرة بأن:

* الخطأ من النوع الأول لإحصائي جودة المطابقة (χ^2_B) و(G^2) تقل بزيادة طول الاختبار ونقصان حجم العينة عند جميع مستويات الدلالة، حيث بلغت الخطأ من النوع الأول عند حجم عينة (١٠٠) مفحوص أقل المعدلات مقارنة بالعينات بحجم (٥٠٠) و(١٠٠٠) مفحوص وكما هو واضح في الجدول عند طول اختبار (٦٠) فقرة وحجم عينة (١٠٠) مفحوص كانت أقل الخطأ للإحصائيين G^2 ، χ^2_B.
* وبشكل معاكس تزداد الخطأ من النوع الأول للإحصائيين χ^2_B ، G^2 بزيادة حجم العينة ونقصان طول الاختبار عند جميع مستويات الدلالة كما هو واضح عند حجم عينة (١٠٠٠) مفحوص وطول اختبار عشرة فقرات.
* ويبدي الإحصائي (RMS) سلوكاً معاكساً حيث تنخفض الخطأ من النوع الأول مع زيادة حجم العينة ونقصان طول الاختبار وتظهر الخطأ من النوع الأول ارتفاعاً عند زيادة طول الاختبار وزيادة حجم العينة، ويتمتع الإحصائي (RMS) بأداء أفضل وبخطأ أقل عند الاختبارات القصيرة عند جميع مستويات الدلالة وعند جميع مستويات حجم العينة مقارنة بأداء الإحصائيين χ^2_B ، G^2 في حين أن أدائهما χ^2_B ، G^2 أفضل عند ظروف ٣٠، ٦٠ فقرة عند جميع مستويات حجم العينة.

- النتائج المتعلقة بالسؤال الخامس "هل تختلف قوة الاختبار الإحصائي لإحصائيات فحص جودة مطابقة الفقرة باختلاف التفاعل بين متغيرات: طول الاختبار وحجم العينة ؟"

يتبين من الجدول (٥):

* أن قوة الاختبار لإحصائي جودة مطابقة الفقرة (G^2) و(χ^2_B) تزداد مع زيادة حجم العينة ونقصان طول الاختبار عند جميع مستويات الدلالة وكما بدت عند طول (١٠) فقرات

وحجم عينة (١٠٠٠) مفحوص حيث أظهر الإحصائي (χ^2_B) قوة إحصائية أكبر من الإحصائي (G^2) لتصل في بعض الظروف إلى واحد.

* في حين أن قوة الاختبار لإحصائي جودة مطابقة الفقرة (RMS) أبدت سلوكاً معاكساً حيث تزداد القوة بزيادة حجم العينة وزيادة طول الاختبار عند جميع مستويات الدلالة الإحصائية كما بدت عند حجم عينة (١٠٠) مفحوص وطول اختبار ٦٠ فقرة.

الفصل الخامس

مناقشة النتائج

الفصل الخامس

مناقشة النتائج

هدفت الدراسة إلى مقارنة ثلاثة إحصائيات لفحص جودة مطابقة الفقرة، تحت ظروف تباين نوع النموذج الرياضي المستخدم، وطول الاختبار، وحجم العينة والتفاعلات بينهما.

وللإجابة على أسئلة الدراسة تم استخدام برمجية (Wingen2) لتوليد بيانات الاستجابة على الفقرات، وبرمجية (Bilog) و(Bilog-mg) لحساب إحصائيات جودة مطابقة الفقرة بهدف المقارنة بينهما تحت ظروف مختلفة من حيث قوة الاختبار والخطأ من النوع الأول.

وحددت مشكلة الدراسة من خلال مجموعة من الأسئلة، وسيتم في هذا الفصل مناقشة وتفسير النتائج المتعلقة بأسئلة الدراسة وتحديد الدرجة التي استطاعت فيها النتائج أن تقدم حلاً لمشكلة الدراسة.

أولا: مناقشة النتائج المتعلقة بالسؤال الأول " هل تختلف الخطأ من النوع الأول وقوة الاختبار الإحصائي لإحصائيات فحص جودة مطابقة الفقرة باختلاف طول الاختبار؟"

*** مناقشة النتائج المتعلقة بأثر طول الاختبار على الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة عند ثبات حجم العينة.**

كشفت النتائج أن الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة (G^2، χ^2_B) تقل مع زيادة طول الاختبار عند جميع مستويات حجم العينة،وتفسر هذه النتيجة بأن الزيادة في طول الاختبار يزيد من الدقة في تقدير القدرة للأفراد مما يجعل تصنيف الأفراد في الفئات الفرعية للقدرة دقيق، ويتم وضع الفرد في الفئة التي تمثل قدرته فعلاً والتوزيع الدقيق للأفراد عبر الفئات الفرعية للقدرة يجعل هنالك اتساق وتطابق ما بين النسب الملاحظة للاستجابات عند كل فئة من الفئات الفرعية للقدرة والنسب المتوقعة والمحسوبة من خلال نموذج استجابة الفقرة، علماً بان نموذج استجابة الفقرة الذي تمت معايرتها عليه البيانات نفس النموذج الذي استخدم في توليد البيانات وفي حساب إحصائي جودة مطابقة الفقرة،ونظرياً فالاختبارات القصيرة تتميز بتقديرات غير دقيقة لقدرات الأفراد الأمر الذي يجعل تصنيف الأفراد في الفئات الفرعية للقدرة غير دقيق ويصنف الفرد في فئة لا تمثل قدرته، مما يؤدي لوجود فرق وعدم تطابق بين النسب الملاحظة للاستجابات عند الفئات الفرعية والنسب المتوقعة والمحسوبة من خلال نموذج استجابة الفقرة والذي يعتمد على وسط أو وسيط القدرات في كل فئة فرعية، والتصنيف الخطأ يؤثر على قيم النسب المتوقعة، وبناء على ذلك فالإحصائيات تشير لعدم المطابقة وتزيد قيمة إحصائي جودة مطابقة الفقرة وتصبح احتمالية الإحصائي ذات

دلالة إحصائية مما يزيد من الخطأ من النوع الأول، وبكلمات أخرى عدم الدقة في تقدير القدرة يولد اختلافاً بين التوزيعات الملاحظة والتوزيعات النظرية لمربع كاي وهذا الانحراف يزيد الخطأ من النوع الأول وتزداد احتمالية رفض الفرضية الصفرية وهي صحيحة.

في حين يبدي إحصائي البواقي المعيارية (RMS) سلوكاً معاكساً بانخفاض الخطأ من النوع الأول بتناقص طول الاختبار ويفسر ذلك بطريقة حساب الإحصائي (RMS) للبواقي إذ أنه يعتمد على تصنيف الأفراد حسب نمط استجاباتهم على الفقرات، إذ يتم تحديد عدد الأنماط الناتجة من استجابات الأفراد على الفقرات،ويصنف الأفراد للفئات الفرعية من خلال النمط الذي أعطاه كل فرد على فقرات الاختبار، حيث أن عدد الأنماط يحدد عدد الفئات الفرعية التي صنف الأفراد إليها، ويحدد بعد ذلك نمط الإجابة على الفقرة عند كل فئة فرعية وبالتالي يحل نمط الإجابة محل تقديرات القدرة في تصنيف الأفراد، وبالتالي فان هذا الإحصائي عالج مشكلة الاختبارات القصيرة في تقديراتها الغير دقيقة للقدرة وذلك باعتماده على نمط الإجابة لتصنيف المفحوصين بدلا من التصنيف اعتمادا على تقديرات القدرة ، ولكنه في حالة الاختبارات الطويلة يكون غير مناسب، فعند الاختبارات القصيرة يكون النمط أدق في التصنيف مقارنة بالاعتماد على القدرة، وخطأ التصنيف أقل، أما في حالة الاختبارات الطويلة فتقديرات القدرة تصبح أكثر دقة وقرار التصنيف المبني عليها أكثر دقة بكثير من الاعتماد على النمط، مما يجعل أداء الإحصائي (RMS) الذي يعتمد على النمط في التصنيف للفئات الفرعية عند الاختبارات الطويلة غير فعال ويتميز بخطأ من النوع الأول مرتفعة.

وبذلك جاءت نتائج هذه الدراسة فيما يتعلق بالإحصائيين G^2 ، χ^2_B تتفق مع نتائج دراسة أورلاندو وثيسن (2000) Orlando and Thissen الذين أكدوا على أن الخطأ من النوع الأول تقل مع زيادة طول الاختبار للإحصائيات ($S-\ \chi^2$) و($S-\ G^2$) وتم فحصها عند ثلاثة مستويات لطول الاختبار (١٠، ٤٠، ٨٠) فقرة. وكذلك دراسة ستون وزانج (2003) Stone and Zhang والتي أشارت إلى أن الخطأ من النوع الأول تقل مع الزيادة في طول الاختبار. ودراسة ستون وهانسن (2000)Stone and Hansen ودراسة أنسلي وبي (1989) Ansley and Bae ودراسة ستون (2003) Stone التي أشارت إلى انحراف توزيع المعاينة لإحصائي جودة المطابقة عن توزيع مربع كاي النظري عندما يكون الاختبار قصير الأمر الذي يرتبط بالتقديرات الغير دقيقة للقدرة، مما يؤدي إلى زيادة الخطأ من النوع الأول في حين أن توزيعات المعاينة لإحصائيات جودة مطابقة الفقرة تقترب من التوزيع

النظري لمربع كاي في الاختبارات الطويلة والتي ترتبط بالتقديرات الدقيقة للقدرة مما يـؤدي إلى انخفاض معدلات الوقوع بالخطأ من النوع الأول.

واتفقت نتائج الدراسة الحالية فيما يتعلق بالإحصائي (RMS) مع نتائج دراسة ستون وزانج (2003) Stone and Zhang والتي أوضحت أن الإحصائيات التي استخدمت التوزيعات البعدية كانت الخطأ من النوع الأول منخفضة وخاصة عند الاختبارات القصيرة وهذا يتوافق مع أداء الإحصائي (RMS) ولكن يختلف أداء الإحصائي (RMS) عن إحصائي ستون χ^{2*} عند زيادة طول الاختبار بسبب اعتماد الإحصائي (RMS) على نمط الإجابة في توزيع الأفراد على الفئات الفرعية للقدرة في حين استخدمت الإحصائيات الأخرى مثل إحصائي ستون χ^{2*} القدرة ولهذا السبب كان أداء هذه الإحصائيات أفضل من أداء الإحصائي (RMS) عند الاختبارات الطويلة.

ب. مناقشة النتائج المتعلقة بأثر طول الاختبار على قوة الاختبار لإحصائيات جودة مطابقة الفقرة عند ثبات حجم العينة.

كشفت النتائج أن قوة الاختبار لإحصائيات جودة مطابقة الفقرة تزداد مع نقصان طول الاختبار وزيادة مستوى الدلالة وتفسير ذلك ففي الاختبارات القصيرة تكون تقديرات القدرة غير دقيقة الأمر الذي يجعل توزيعات المعاينة لإحصائيات جودة مطابقة الفقرة تنحرف وتختلف عن التوزيع النظري لمربع كاي بسبب أن التوزيع الغير دقيق للأفراد عبر الفئات الفرعية للقدرة يجعل هنالك اختلاف ما بين النسب الملاحظة والمتوقعة عند كل فئة من الفئات الفرعية للقدرة ما يجعل فرصة قرار رفض الفرضية الصفرية أكبر وخاصة للإحصائي G^2 ، χ^2_B في حين انخفضت القوة للإحصائي (RMS) عند الاختبارات القصيرة (١٠) فقرات وذلك لأنه تجاوز مشكلة الاختبارات القصيرة المتعلقة بالتقديرات الغير دقيقة للقدرة واعتماده على نمط الإجابة والتوزيعات البعدية. ما جعل فرصة رفضه للفرضية الصفرية في هذا الظرف أقل ولكنها عادت بالارتفاع عند الأطوال ٣٠، ٦٠ فقرة لأن اعتماده على نمط الإجابة في الاختبارات الطويلة أسلوب غير دقيق في تصنيف المفحوصين ويحدث أخطاء في التصنيف مما يجعل التوزيعات المعاينة تختلف عن التوزيعات النظرية وتزداد فرصة رفض الفرضية الصفرية وتزداد قوته عند الاختبارات الطويلة.

وجاءت نتائج هذه الدراسة تتفق مع دراسة ستون وهانسن (2003) Stone and Hansen والذين أكدوا على أن قوة الاختبار لإحصائيات جودة مطابقة الفقرة تقل عندما تقل α وتزداد عندما يقل طول الاختبار.

<u>ثانياً:</u> مناقشة النتائج المتعلقة بالسؤال الثاني" هل تختلف الخطأ من النوع الأول وقوة الاختبار الإحصائي لإحصائيات لفحص جودة مطابقة الفقرة باختلاف حجم العينة؟."

***. مناقشة النتائج المتعلقة بأثر حجم العينة على الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة عند ثبات طول الاختبار.**

كشفت النتائج أن الخطأ من النوع الأول لإحصائيات جودة مطابقة الفقرة G^2،χ^2_B تزداد بزيادة حجم العينة عند ثبات طول الاختبار وتفسر هذه النتيجة بأن الزيادة في حجم العينة يزيد عدد الأفراد الواقعين في كل فئة فرعية وعندما تترافق هذه الزيادة مع تصنيف غير دقيق للأفراد عبر فئات القدرة، الأمر الذي يزيد من انحراف التوزيعات الملاحظة (توزيعات المعاينة) عن التوزيع النظري لمربع كاي ويمكن ملاحظة ذلك عند طول اختبار عشرة فقرات حيث ارتفعت الخطأ من (٠,١٨٢) عند حجم عينة (١٠٠) مفحوص إلى (٠,٨٦) عند حجم عينة (١٠٠٠) مفحوص بالنسبة للإحصائي G كما في الجدول (٤).

وكشفت النتائج بأن الإحصائي (χ^2_B) يتمتع بخطأ من النوع الأول أقل مقارنة بالإحصائي (G^2) عند حجم عينة (١٠٠) مفحوص في حين أن الإحصائي (G^2) يتمتع بخطأ من النوع الأول أقل عند مستويات حجم العينة (٥٠٠) و(١٠٠٠) مفحوص مقارنة بالإحصائي (χ^2_B).

وتفسير هذه النتيجة عند تفحص النماذج الرياضية للإحصائيين كما وضح في الإطار النظري للدراسة أن الإحصائي بك (χ^2_B) يعتمد في حسابه للنسب المتوقعة للفئات الفرعية للقدرة على وسيط قدرة المفحوصين الواقعين في كل فئة فرعية في حين أن الإحصائي (G^2) يعتمد في حسابه للنسب المتوقعة للفئات الفرعية للقدرة على وسط قدرة المفحوصين الواقعين في كل فئة فرعية، ونجد في حالة العينات ذات الأحجام الصغيرة، أن عدد الأفراد الواقعين في كل فئة فرعية قليل، مما يجعل استخدام الوسيط لقدرات المفحوصين الواقعين في الفئة الفرعية أكثر ملائمة من الوسط لحساب النسب المتوقعة وذلك لعدم تأثر الوسيط بالقيم الكبيرة والتقديرات الغير دقيقة للقدرة والتي تؤثر في قيمة القدرة التي تدخل في حساب النسبة المتوقعة عند تلك الفئة.

في حين أن استخدام الوسط في حالة العينات الكبيرة (٥٠٠، ١٠٠٠) مفحوص أكثر ملائمة لأن عدد الأفراد الواقعين في الفئات الفرعية سوف يزيد، وتأثير القيم الشاذة سيقل عند زيادة حجم العينة ويجعل قيمة القدرة التي تدخل في حساب النسبة المتوقعة أكثر دقة وفاعلية في تحديد النسبة المتوقعة عند كل فئة فرعية، إضافة لذلك استفاد الإحصائي (G^2) من

استخدام صيغة الأرجحية العظمى في تحسين أدائه في حين اعتمد الإحصائي χ^2_B على الصيغ العادية لبيرسون.

وجاءت نتائج الدراسة تتفق مع نتائج دراسة هامبلتون وآخـرون Hambleton et al. (1991) التي أشارت إلى زيادة عدد الفقرات غير المطابقة مع زيادة حجـم العينـة ودراسات التي أكدت على زيادة الخطأ من النوع الأول بزيادة حجـم منهـا (Orlando and Thissen ,2000; Stone and Hansen ,2000; Stone and Hansen ,2000; Stone ,2003)

في حين يبدي إحصائي البواقي المعيارية (RMS) معدلات منخفضة في الخطأ مـن النوع الأول مقارنة بالإحصائيين χ^2_B ، G^2عند طول عشرة فقرات وعند جميع مستويات حجم العينة، ويرجع ذلك لطريقة حسابه والمتعلقة بتوزيع الأفراد على الفئات من خلال نمط الإجابة والذي يجعل توزيعات المعاينة (الملاحظة) مطابقة للتوزيعات المتوقعة (التوزيعات النظريـة) لمربع كاي بالإضافة لوجود اقتران التوزيعات البعدية الذي يدخل في حساب قيمة الإحـصائي ولكنه يتميز بارتفاع الخطأ من النوع الأول في الاختبارات الطويلة (٣٠) و(٦٠) فقرة وعنـد جميع مستويات حجم العينة مقارنة بالإحصائيين G^2 ، χ^2_B ،لأنه في حالة الاختبارات الطويلة يكون غير مناسب كما وضح سابقا، واتفقت النتائج مع نتائج دراسـة سـتون وزانـج Stone and Zhang (2003) بانخفاض الخطأ من النوع الأول عند الاختبارات القصيرة عند استخدام إحصائي ستون الذي يعتمد على التوقعـات البعديـة ويختلـف مـع (RMS) عنـد الاختبارات الطويلة لأن الإحصائي ستون χ^2 يعتمد على تقديرات القدرة بينمـا الإحـصائي (RMS) يعتمد على نمط الإجابة في تصنيف الأفراد على فئات القدرة الفرعية.

***مناقشة النتائج المتعلقة بأثر حجم العينة على قوة الاختبار لإحصائيات جودة مطابقة الفقرة عند ثبات طول الاختبار.**

كشفت النتائج أن قوة الاختبار لإحصائيات جودة مطابقة الفقرة تزداد مع الزيادة فـي حجم العينة وزيادة مستوى الدلالة وتفسير ذلك ففي العينات الكبيرة والمترافقة بالتقديرات الغير دقيقة للقدرة الناتجة عن استخدام الاختبارات القصيرة يجعل توزيعات المعاينـة (التجريبيـة) (الملاحظة) بإحصائيات جودة مطابقة الفقرة تنحرف عن التوزيعات النظرية لمربع كاي ممـا يجعل احتمال رفض الفرضية الصفرية أكبر وخاصة للإحصائيين G^2 ، χ^2_B.

في حين انخفضت القوة الإحصائية للبواقي المعياريـة (RMS) عنـد الاختبـارات القصيرة بسبب اقتراب التوزيع التجريبي (المعاينة) من التوزيع النظري لمربع كاي، بـسبب

استخدامه للتوقعات البعدية في حين تميز بقوة أكبر مقارنـة بالإحـصائيين الـسابقين عنـد الاختبارات ذات طول (٣٠) و(٦٠) فقرة وعند جميع مستويات حجم العينة بسبب عدم دقتـه والناتجة عن تصنيف الأفراد حسب نمط إجابتهم التي لا تتناسب والاختبارات الطويلة، الأمـر الذي جعل احتمالية رفض الفرضية الصفرية كبيرة. واتفقت مع نتائج الدراسات التـي أكـدت على أن قوة الاختبار تزداد بزيادة حجم العينة ومنها Stone and Zhang ,2003; Farish; (1984, Stone ,2003).

ونظرياً احتمال الوقوع بالخطأ من النوع الأول يدعى (α) واحتمال الوقع بالخطأ مـن النوع الثاني (β) وقوة الاختبار الإحصائي بشكل عام تزداد بنقصان قيمة β وزيـادة قيمـة α ولكنها ليست البديل الواقعي ولزيادة قوة الاختبار لا بد من زيادة حجم العينة، لأن لحجم العينة علاقة بتوزيع المجتمع فكلما زاد حجم العينة كلما كان هنالك احتمال أقل لاقتـراب التوزيـع التجريبي الذي يكون مرتبط بالتوزيعات الغير دقيقة للأفراد عبر الفئات المختلفـة بـالتوزيع النظري وبالتالي القرار المتعلق برفض الفرضية الصفرية عندما تكون خاطئة يكون أقوى.

ثالثاً: مناقشة النتائج المتعلقة بالسؤال الثالث" هل تختلف الخطأ من النوع الأول لإحصائيات فحص جودة مطابقة الفقرة باختلاف النموذج الرياضي المستخدم ؟."

كشفت النتائج أن الخطأ من النوع الأول لإحصائيات جـودة مطابقـة الفقـرة تحـت افتراض النموذج ثلاثي المعلم بشكل عام وفي معظم الظروف الناتجة عن الخلايا البحثية فـي التصميم أعلى بقليل من الخطأ من النوع الأول تحت افتراض النموذج ثنائي المعلم.

ونظرياً قيم إحصائيات جودة مطابقة الفقرة تتأثر بمعلمتي التمييز والتخمين ولا تتـأثر بمعلمة الصعوبة، فبزيادة قيم التمييز والتخمين تزيد قيم إحـصائيات جـودة مطابقـة الفقـرة وبالتالي زيادة احتمالية رفض الفرضية الصفرية الذي ينتج زيادة في الخطأ من النـوع الأول، ولأن الفقرات تم توليدها تحت النموذج ثنائي المعلم في الجدول (٣) وثلاثي المعلم في الجدول (٤) تتشابه في توزيعات الصعوبة والتمييز، والاختلاف الوحيد هو وجود معلمة التخمين فـي النموذج الثلاثي الأمر الذي جعل عدد الفقرات غير المطابقة باستخدام النموذج الثلاثي أكبـر في الجدول (٤) والتي حسبت فيها إحصائيات جودة مطابقة الفقرة. واتفقت نتائج هذه الدراسة مع نتائج دراسات الت بأنه لا توجد علاقة ما بين قيم إحصائيات مطابقة الفقـرة مـع معلـم الصعوبة وان زيادة مستوى تمييز الفقرة أو التخمين يؤدي لزيادة في قيم إحصائيات مطابقـة الفقرة الأمر الذي يؤدي لزيادة الخطأ مـن النـوع الأول ومنهـا (Dodeen2004; Seol ,1999; Movnt and Schumacker ,1998)

رابعاً: مناقشة النتائج المتعلقة بالسؤال الرابع "هل تختلف الخطأ من النوع الأول لإحصائيات فحص جودة مطابقة الفقرة باختلاف التفاعل بين متغيرات: طول الاختبار وحجم العينة؟".

كشفت النتائج بأن الخطأ من النوع الأول لإحصائي جودة مطابقـة الفقـرة G^2 ، χ^2_B تقل بزيادة طول الاختبار ونقصان حجم العينة وتفسير ذلك أنه بزيادة طول الاختبار ونقصان حجم العينة يجعل التوزيعات التجريبية (المعاينة) تقترب مع التوزيع النظري لمربـع كـاي. الأمر الذي يقلل من احتمالية الوقوع بالخطأ من النوع الأول واتفقت النتائج مع نتائج دراسـة ستون وزانج Stone and Zhang (2003) الذين أكدوا أنه بزيادة طول الاختبار ونقـصان حجم العينة تقل الخطأ من النوع الأول.

في حين أن إحصائي البواقي المعيارية (RMS) تنخفض الخطـأ مـن النـوع الأول بزيادة حجم العينة ونقصان طول الاختبار وذلك بسبب اعتماده على نمط الإجابة في توزيـع الأفراد إضافة لاعتماده على التوقعات البعدية كما اتضح في بنائه الرياضي، وجاءت النتـائج متفقة مع دراسة ستون وزانج فيما يتعلق بالإحصائي χ^2 والذي* يعتمد على التوقعات البعديـة في حسابه عند طول (١٠) فقرات. واختلف معه عند الاختبارات الطويلة وذلك بسبب اعتمـاد إحصائي ستون على القدرة في توزيع المفحوصين على الفئات والتي تكون أكثـر دقـة مـن استخدام نمط الإجابة كما في الإحصائي (RMS).

خامساً: مناقشة النتائج المتعلقة بالسؤال الخامس "هل تختلـف قـوة الاختبـار الإحـصائي لإحصائيات فحص جودة مطابقة الفقرة باختلاف التفاعل بين متغيرات: طـول الاختبـار وحجم العينة ؟"

كشفت النتائج أن قوة الاختبار لإحصائيات جودة مطابقة الفقرة G^2 ، χ^2_B تزداد بزيادة حجم العينة ونقصان طول الاختبار عند جميع مستويات الدلالة وذلك لأن زيادة حجم العينـة ونقصان طول الاختبار تجعل التوزيعات التجريبية (المعاينة) لإحصائي جودة مطابقة الفقـرة G^2 ، χ^2_B تنحرف عن التوزيع النظري لمربع كاي مما يجعل احتماليـة رفـض الفرضـية الصفرية كبير، والناتج عن زيادة قيم إحصائيات جودة مطابقـة الفقـرة بـسبب انحرافـات التوزيعات النظرية والمعاينة عن بعضها البعض. وهذا يتفق مع دراسة ستون وزانج Stone and Zhang(2003) الذين أكدوا على زيادة قوة الاختبار بنقصان طول الاختبار وزيادة حجم العينة.

في حين أن إحصائي البواقي المعيارية (RMS) أبدى معدلات قوة مرتفعة عند زيادة حجم العينة وزيادة طول الاختبار، كما وضح سابقاً وتحت هذه الظروف تنحرف التوزيعـات

المعاينة للإحصائي مطابقة الفقرة (RMS) عن التوزيعات النظرية للإحصائي مما يزيد مـن احتمالية رفض الفرضية الصفرية وعدم اتساق التوزيع النظري مع توزيع المعاينة، واتفقـت نتائج الدراسة مع نتائج دراسة ستون وزانج أيضا الذين أكدوا على زيادة قوة الاختبار بزيـادة حجم العينة وزيادة طول الاختبار للإحصائيات المرتبطة بالتوقعات البعدية.

التوصيات و الاستنتاجات:

من خلال النتائج التي تم التوصل إليها يوصي الباحث بما يلي:

- إجراء دراسات توضح أثر الطريقة المستخدمة في تقدير القدرة على الخطأ من النوع الأول وقوة الاختبار لإحصائيات جودة مطابقة الفقرة .
- إجراء دراسات توضح أثر تباين عدد الفئات الفرعية للقدرة في الخطأ مـن النـوع الأول وقوة الاختبار لإحصائيات جودة مطابقة الفقرة.
- إجراء دراسات توضح أثر عدد مرات التدوير(Iterations or Cycles) المـستخدم في تقدير القدرة على الخطأ من النوع الأول وقوة الاختبار لإحصائيات جودة مطابقة الفقرة.
- إجراء دراسات توضح أثر التوزيعات الغير سوية للقدرة و صعوبة الفقرة على الخطأ من النوع الأول وقوة الاختبار لإحصائيات جودة مطابقة الفقرة.

المراجـــع

أولا: المراجع العربية

علام، صلاح الدين محمود.(١٩٨٦).**تطورات معاصرة في القياس النفـسي و التربـوي**. الكويت:مطابع القبس التجارية .

ثانياً: المراجع الأجنبية:

Ankenmann, R. D. (1994/1995). Goodness Of Fit And Ability Estimation In The Graded Response Model**.**(Doctoral dissertation, university of Pittsburgh, 1994). **Dissertation abstracts international**, 55 (10), 3167.

Ankenmann, R. D., and Stone, C.A. (1992, April). A Monte Carlo study Of marginal maximum likelihood estimates for the graded model. Paper presented at annual meeting of the national council of measurement in Education**,** San Francisco, CA. (**ERIC Document Reproduction Services** NO ED. 347208).

Ansley, T.N., and Bae, H.W (1989, April). An Empirical investigation of the nature of the distribution of an IRT goodness –of– fit statistic. **Paper Presented At The Annual Meeting Of The American Educational Research Association, San Francisco, CA**.

Baghi, H. (1990). The Use of rasch model fit Statistics In Selecting Items For The Maryland Functional Testing Program**. (ERIC Document Reproduction Service** NO. ED324363**).**

Crocker,L.,andAlgina,J.(1986).**Introduction To Classical And Modern Test Theory** .Holt Rinehart and wiston , New YORK.

Dodeen, H. (2004).The relationship between item parameters and item fit. **Journal Of Educational Measurement**. 40 (3), 261-270.

Farish,S.(1984).Investigating item stability: An empirical investigation in to the variability of item statistics under conditions of varying sample design and sample size. **(ERIC Document Reproduction Service** NO. ED262046**).**

Hambelton, R.K. (1993). Principles and selected applications of item response theory**.** In R.L. Linn (Ed). **Educational Measurement**. (3rd ed.) (pp.147-200). Phoenix: the oryx press.

Hambleton, R.K.,and swaminathan, H. (1985). **Item Response Theory: Principles And Applications.** Boston, MA: Kluwer- Nijhoff.

Hambleton,R.K,swaminathan,H.,andRogers, Jane H. (1991). **Fundamental Of Item Response Theory.** Newbury park, CA: sage publications, Inc.

Han,K.T.,andHambelton,R.K.(2007).**UserManualForWinGen:Windons Software That Generates IRT Model Parameters And Item Response.** Holt Rinehart and wiston , New YORK.

Hansen, M. A. (2004). **Predicting The Distribution Of A goodness – Of- Fit Statistics Appropriate For Use With Performance – Based Assessments** (Doctoral dissertation, university of Pittsburgh, 2004).

Harwell,M.et al.(1996). Monte Carlo Studies in Item Response Theory .**Applied Psychological Measurement**,vol.20,No.2.

Lord, F. M. (1990). **Application of item response theory to practical testing problems.** Hill sdale, new Jersey 07642.

Mckinley, R. L., and Mills, C. N. (1985). A comparison of several goodness –of- fit statistics**. Applied psychological measurement** 9(1), 49-57.

Mote, V. L., and Anderson, R. L. (1965). An investigation of the effect of misclassification on the properties of chi-square tests in the analysis of categorical data**. Biometrika**, 52, 95-110.

Movnt, R. E., and schumacker, R.E. (1998). Identifying measurement disturbance effects using Rasch item fit statistics and the logit residual index**. Journal Of Out Comes Measurement**, 2 (4), 338-350.

Orlando, M., and Thissen, D. (2000). Likelihood- based item- fit indices for dichotomous item response theory models**.Applied Psychological Measurement**, 24(1), 50-64.

Reise, S.P. (1990). A comparison of item –fit and person- fit methods of assessin model- data fit in IRT**. Applied psychological measurement.** 14(2), 127-137.

Rogers, H. J., Hattie, J. A. (1987). A monte carlo investigation of several person and item fit statistics for item response models. **Applied Psychological Measurement**, 11(1), 47-57.

Seol, H. (1999). Detecting differential item functioning with five standardized item –fit indices in the Rasch model. **Journal Of Outcomes Measurement**, 3(3), 233-247.

Smith, R.M. (1991). The distributional properties of Rasch item fit statistic. **Educational And Psychological Measurement**, 51, 541-565.

Stone, C. A. and Hansen, M. A. (2000). The effect of errors in estimating ability of goodness –of- fit tests for IRT models. **Educational And Psychological Measurement**, 60(6), 974-991.

Stone, C. A. and Zhang, B. (2003). Assessing goodness –of- fit of IRT models. A comparison of traditional and alternative procedures. **Journal Of Educational Measurement**, 40(4), 331-352.

Stone, C. A., Ankenmann, R. D., Lane, S., and Liue, M. (1995). Examination of the assumptions and properties of graded item response theory model, An example using a mathematics performance assessment. **Applied Measurement In Education,** 8(4), 313-340.

Stone, C. A., Mislevy, R.J., Mazzeo, J. (1994, April). Classification error and goodness –of- fit in IRT models**. Paper presented at the annual meeting of the American educational research association,** new Orleans**.**

Tiot,M.D.(2003).**IRT From SSI: BILOG-MG MULTILOG PARSCALE TESTFACT.**Scientific Software International.Inc.

Wright, B. D. and Panchapakesan, N. (1969). A procedure for sample free item analysis**. Educational And Psychological Measurement**, 51, 541-565.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. **Applied Psychological Measurement**, 5, 245-262.

الملحق ١. الأوامر الخاصة بحساب إحصائي جودة مطابقة الفقرة(بك , χ^2_B)

This syntax will compute the chi-square item fit statistics, (BOCK'S INDEX fit statistics index to measure the goodness of fit of the model t each item in the test for the data set.

```
>COMMENTS;

   This is syntax of the BILOG Calibration of test.
 the chi-square item fit statistics, (BOCK'S   INDEX)



>GLOBAL  NPARM=--- ,DFName = 'D:\onethuosand60item
power_100.wgr',SAVE';
>SAVE    SORT='BLGDAT\FIT.SOR',GRAPH='FIT.PLT';
>LENGTH  NITEMS=60;
>INPUT  NALT=2,NIDCH=8,SAMPLE=1000, CODE='01';
        (8A1,2X,60A1)
>TEST    TNAME=RAJI;
>CALIB   CYCLES=20,NEWTON=2,CHISQR=60;
>SCORE   METHOD=2;
```

الملحق ٢. الأوامر الخاصة بحساب إحصائي جودة مطابقة الفقرة(RMS) البواقي المعيارية

This syntax will compute the (Mean-square standardized deviates replace the chi-square item)(RSM) fit statistics index to measure the goodness of fit of the model t each item in the test for the data set.

```
>COMMENTS;

Mean-square standardized deviates replace the chi-square item fit statistics.



>GLOBAL  NPARM=3 DFNAME='BLGDAT\FIT.DAT',SAVE';
>SAVE    SORT='BLGDAT\FIT.SOR',GRAPH='FIT.PLT';
>LENGTH  NITEMS=60;
>INPUT  NALT=2,NIDCH=8,SAMPLE=100, CODE='01';
       (8A1,2X,60A1)
>TEST    TNAME=RAJI;
>CALIB   CYCLES=20,NEWTON=2,POST;
>SCORE  METHOD=2;
```

الملحق ٣. الأوامر الخاصة بحساب إحصائي جودة مطابقة الفقرة(الارجحية العظمى ,G^2)

This syntax will compute the G-SQUARE index to measure the goodness of fit of the model t each item in the test for the data set.

```
>GLOBAL DFName = 'C:\Documents and Settings\a\ سطح
المكتب\onethuosand60item  power_100_1.wgr',
        NPArm = 3,
        LOGistic;
>LENGTH NITems = (60);
>INPUT NTOtal = 60,
       NALt = 1000,
       NIDchar = 8;
>ITEMS ;
>TEST1 TNAme = 'TEST0001',
       INUmber = (1(1)60);
(8A1, 2X, 60A1)
>CALIB ACCel = 1.0000,
       CHIsquare = (60, 9);
>SCORE ;
```

COMPARISON OF THREE GOODNESS OF FIT STATISTICS UNDER CONDITIONS OF TYPES OF LOGISTIC MODELS, TEST LENGTH, SAMPLE SIZE AND IT'S INTERACTIONS .

By

Raji A. Saraierh

Supervisor

Dr. Kalial Elean, Prof.

ABSTRACT

This study aim to Comparison of three goodness of fit statistics under conditions of types of logistic models, test length, sample size and it's interactions on the type I error rates and empirical power, for the goodness of item fit strategies. To prove the aim of this study, the researcher used simulation method to generate dichotomous response on the response items by using the software application Wingen 2. The researcher divided the sample to a three different groups (100,500,1000) examinee and three different length of test (10,30,60) Item, to investigate type I error rates, the percent of misfit detected across the 100 replications was calculated for those items in which $H0$ was true (data were simulated under a 2-P model and goodness of fit was assessed using item parameter estimates from a 2-P model or data were simulated under a 3-P model and goodness of fit was assessed using item parameter estimates from a 3-P model). Three different α levels were selected to examine the behavior of the goodness-of-fit tests at (0.1 ,0.05, 0.01).

To investigate empirical power, the percent of misfit detected across 100 replications for each combination of test length and sample

size was calculated given α = 0.1, .05, and .01 for items in which ***H0*** was false(data were simulated under a 3-P model and goodness of fit was assessed using item parameter estimates from a 2-P model).

The result of study as follows:

- Decreased Type I error rates for goodness of fit strategy χ^2_B, G^2 with increased the test length and Decreased sample size, while the decreased type I error rates for goodness of fit strategy Root-Mean-Square of the posterior deviates (RMS) with increased the sample size and decrease the test length.
- Increased empirical power for goodness of fit strategy χ^2_B, G^2 with increased the sample size and decrease the test length, while increased empirical power for goodness of fit strategy root-mean-square of the posterior deviates (RMS) with increased the sample size and increased the test length .
- Type I error rates for goodness of fit strategies χ^2_B, G^2 , (RMS) under of assumptions 3-P model higher than Type I error rates under of assumptions 2-P model for all conditions.